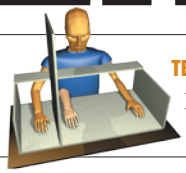
THIS WEEK

EDITORIALS

HELIUM Ancient stocks of noble gas escape from Yellowstone **p.266**

WORLD VIEW The patent nonsense of WHO disease projects **p.267**



TENDER Soft skin strokes put rubber hand on a limb **p.268**

Not so neutral

Switzerland's science landscape is under threat after a narrow majority of citizens voted for tighter immigration rules that could restrict the number of foreign scientists who work in the country.

Its great desire for independence notwithstanding, Switzerland is superbly integrated into the global pursuit of scientific research. The small Alpine country maintains some of Europe's strongest research universities and hosts one of the world's most outstanding research facilities, CERN — Europe's particle-physics laboratory near Geneva. The nation has also taken on a lead role, financially and logistically, in the Human Brain Project, a €1-billion (US$1.4-billion) collaboration to simulate the human brain in a supercomputer.

The outcome of an ill-conceived referendum on 9 February against 'mass immigration' threatens to spoil Switzerland's beautiful science landscape (see page 277). The motion was approved by a narrow majority despite opposition from the government, parliament and Swiss lobby groups, including those from science and industry.

The government now has three years to implement tighter immigration rules that could restrict how many foreign scientists can be employed at the country's universities and research institutes. The European Union (EU) has already shown that it is not going to relax its fundamental principles in deference to Swiss xenophobia. Last weekend, Switzerland — an EU associate member — refused to sign an agreement that would extend the right to free movement within the EU to the bloc's newest member, Croatia. The European Commission immediately suspended talks over Switzerland's association in the EU's €80-billion Horizon 2020 research programme. Clearly, the nation's U-turn will not be without consequences.

Switzerland prides itself on a tradition that grants its citizens more far-reaching rights of co-determination than any other democracy. Its relatively small geographical size and population, high level of literacy and well-developed pragmatism all seem to favour that special form of government.

But direct democracy becomes problematic if it is driven by populism and irrational fears, such as those over unemployment and crime (Switzerland is, in fact, one of the safest countries in the world, and the current unemployment rate is barely 3.5%). Certainly, immigration there has increased over the past decade — but this is in large part because the economy and health system rely heavily on the services of foreign workers. Ironically, the initiative to 'stop mass immigration' got the highest level of support in rural areas, where there are relatively few foreigners. In cosmopolitan cities, such as Zurich, Basle and Geneva, a majority of voters rejected the initiative.

This is not the first time that an initiative by the Swiss people has clashed with the interests of science. In 1992, they voted to include the protection of the dignity of animals in the constitution, which made animal experiments much harder for scientists to justify. In 1998, an initiative aimed at banning the use of transgenic animals failed — narrowly — after scientists lobbied vociferously against it.

Fortunately, the wording of the latest initiative offers enough leeway for the Swiss government to avert unintended harm to science. For example, the government is free to include provisions that would exempt foreign scientists — and possibly other groups of professionals, such as nurses — from the restrictions altogether. It needs to do so. Alternatively, it might assign future immigration quotas regionally, so that rural cantons could restrict immigration more than the urban regions that host universities and research institutes, and which voted against the measure.

The European Commission should stand firm on its decision to halt Horizon 2020 talks. It is unfortunate that science will be a casualty of a broader political fight but, in this case, principles matter. Switzerland cannot play fast and loose with international agreements such as those on the free movement of Europeans. If the country is to remain at the forefront of prestigious international research collaborations such as the Human Brain Project, it must make assurances that scientists can continue to participate on Swiss soil, and find a way to make it happen.

"Switzerland cannot play fast and loose with international agreements on free movement."

Switzerland's regrettable course is a setback to Brussels' vision of a pan-European Research Area where scientists, knowledge and ideas can move freely across borders. Indeed, it is a setback to any attempt to fight the populist rhetoric that immigration is a threat.

As the drama plays out between Bern and Brussels, European scientists and science organizations should seek to maintain, and where possible enhance, mutual collaboration with this exceptionally science-minded nation. But in this centenary year of the outbreak of the First World War — modern Europe's original sin — Switzerland must be reminded that nationalism and exclusion are anachronisms of the worst kind. ■

Intelligent testing

Science has a part to play in ensuring protection for defendants with intellectual disabilities.

Most societies recognize that it would be wrong to execute someone like Lennie Small, the mentally disabled character in John Steinbeck's 1937 novella *Of Mice and Men*, even though he murdered a woman while stroking her soft hair.

Lennie never understood what he did wrong, and that ignorance usually brings protection from the full force of the law. Most countries with a death penalty have some sort of special treatment for the mentally disabled enshrined in the judicial system.

Few intellectually disabled people are as obviously impaired as Lennie, so expert assessments — and science — are usually used to

decide their fate. Most US states, for instance, use an IQ test to assess cognitive skills such as problem-solving and anticipating the consequences of actions. Lennie would have scored low: when the woman screamed, his reasoning power was so limited that he could come up with no other option but to kill her. The tests can provide an accurate measure of some cognitive skills and, better yet, seem to offer an objective metric for prosecutors to work with. No test of cognitive ability, however, can determine a person's understanding of guilt, and thus their culpability for a crime. That problem becomes especially difficult when defendants have a mild intellectual disability.

Faced with such cases, some court systems use IQ score as a proxy to assess the deeper issue of awareness. Florida is one, and early next month its controversial approach will be tested. As we report on page 284, on 3 March the US Supreme Court will begin hearing arguments on behalf of Freddie Lee Hall, a convicted murderer. Hall has a low IQ, but not consistently low enough (below 70) to escape the death penalty in Florida. He and his lawyers want the state to raise its IQ cut-off point.

The state has refused. A weakening of its criteria, officials say, could prompt hundreds of appeals. One estimate suggests that in the United States, up to 20% of the 3,100 or so people on death row may have some level of intellectual disability (R. Coyne and L. Entzeroth *Geo. J. Fighting Pov.* **3,** 40; 1996). And if it relaxes its strict interpretation, the state worries, could not a clever lawyer or sympathetic psychiatrist claim that a client facing the death penalty has a mental disability due to post-traumatic stress disorder, temporary insanity or a bout of depression? Defendants, lawyers and officials in other states are watching with interest.

If the United States is to have a death penalty — and 55% of Americans supported it in a 2013 survey — it should ensure that all defendants have an equal, objective chance to save their own lives. Many states try to ensure this by drawing a 'bright line' at an IQ score of 70. But this greatly overestimates the IQ test's precision. The tests have a ten-point margin of error — they cannot necessarily distinguish a 71 from a 69. If IQ tests were to be scrapped as a way to judge criminal competence, what could replace them?

In its latest version of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*), the American Psychiatric Association changed both the definition and the name of intellectual disability, formerly known as mental retardation. It now avoids setting any IQ limit for the disorder, and emphasizes the impact of cognitive ability on behaviour.

Related to this approach is the adaptive behaviour test. Designed to measure how well a person can manage in the real world by quizzing his or her family and acquaintances, when administered by experts, this standardized test gives the kind of consistent, numerical results that prosecutors crave.

"Too little is known about how intelligence plays into criminality."

Psychologists in the United States are already designing a modified version called the Diagnostic Adaptive Behavior Scale, the first evidence-based, adaptive behaviour test designed specifically for young people with a low IQ. Relevant to the debate over mental dysfunction and the death penalty, it assesses traits such as gullibility and the ability to solve social problems. Properly administered, it could determine awareness for courts better than existing tests of IQ.

Too little is known about how intelligence plays into criminality and the various environmental factors that affect it (such as decades spent in prison). If science is to provide courts with more certainty about the state of mind of defendants, then more research is needed on the nature of intelligence itself.

For example, Kent Kiehl, a psychologist at the University of New Mexico in Albuquerque, is compiling 3,000 brain scans of prisoners in what is already the largest collection of images of criminal minds in the world. Kiehl's main aim is to assess factors such as psychopathy, or whatever it is that makes people commit crimes, but he also wants to develop a test to predict intelligence. Adding this to the arsenal of such tests could help to assuage prosecutors' concerns about a defendant faking disability, or an expert giving a biased diagnosis.

In Steinbeck's book, Lennie pays the ultimate price for both his crime and his disability. Justice demands that we separate the two. Science will keep trying to do so. ■

Helium high

Many bemoan the shortage of helium for the lab, but for geologists, its true value is in the ground.

The ancient mariner and his ill-fated shipmates in Samuel Taylor Coleridge's epic poem were tormented by the sight of "Water, water, every where/Nor any drop to drink". That sentiment is likely to be shared by physicists and other researchers who have struggled in recent years to find cheap helium for their studies, equipment and experiments, when they read that massive quantities of the gas have been found escaping from the well-trodden and turbulent ground of Yellowstone National Park in Wyoming.

And we do mean massive amounts. Perhaps one billion years' worth of stored helium is fizzing up from Earth's crust beneath Yellowstone, only to disappear into thin air.

Meanwhile, the US Geological Survey (USGS) reported last month that helium prices reached an 18-year high in the 12 months to September 2013 — around the same time that the US Congress voted to postpone closure of the nation's strategic helium reserve. In doing so, Congress overruled an older law requiring that the United States sell off supplies of the gas that it has hoarded since the 1920s — an economic albatross around the neck of the laws of supply and demand that many blame for the current price volatility (see *Nature* http://doi.org/rkc; 2013).

The United States is by some distance the world's largest helium supplier. Yet there is unlikely to be a sensible and affordable way to tap the gas flooding into the atmosphere at Yellowstone. So, the waste is inevitable. Instead of bemoaning it, admire the science it brings. As useful as extracted, processed and packaged helium is to researchers in the lab, the true value of this noble gas for Earth scientists lies in the ground. Subterranean helium is a crucial geophysical tracer and one used, for example, to date groundwater and to track the rise of the continents.

Around the world, geysers and hot springs bubble this telltale helium to the surface. The ratios of helium isotopes in such escapes provide clues about the characteristics of volcanic activity in the crust and mantle. At Yellowstone, these isotopes help geologists to make sense of a particular super-volcano feature called the Yellowstone hotspot.

Much of the helium emitted at Yellowstone is helium-4, an isotope produced by radioactive decay of elements such as uranium and thorium in the crust. (The other common isotope, helium-3, is a primordial relic of the formation of the planet.)

On page 355 of this issue, Jacob Lowenstern and his colleagues at the USGS show that the helium-4 emission rates from Yellowstone exceed any conceivable rate of generation within the crust. Instead, it must have accumulated in the crust underneath Yellowstone for hundreds of millions of years, until the geological carnage of the Yellowstone hotspot allowed it to escape. Perhaps one billion years' worth has been liberated over the past two million years. And still it comes. As Coleridge said, "The very deep did rot: O Christ!/That ever this should be!" ■

➲ NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv


POLICY CURES

WHO plans for neglected diseases are wrong

Research and development into diseases affecting the world's poorest people will not benefit from the agency's policy, warns **Mary Moran**.

After more than a decade of trying to find a way to fund research on diseases that affect the developing world, the World Health Organization (WHO) made a decisive move last month when it announced its first pilot projects. As *Nature* reported (see *Nature* **505,** 142; 2014), the WHO hopes that these projects will break the stalemate over research on neglected conditions such as kala-azar, a deadly parasitic disease that afflicts hundreds of thousands of the world's poorest people.

The WHO is taking giant strides, but they are in the wrong direction. The projects are based on flawed logic and will waste time and money. Worse, this initiative could actively damage existing projects to develop such medicines. The WHO pilot should be stopped.

I do not make these claims lightly. I was involved in the WHO analysis, drafting and recommendations, and know how difficult it has been.

The pilot projects are the culmination of a ten-year negotiation that aimed to achieve two goals: to make commercial medicines more affordable for the developing world, and to stimulate public (non-profit) development of medicines for neglected diseases.

The first goal demands complex and contentious efforts to change or replace the commercial pharmaceutical system, which funds its drug research and development (R&D) programmes by charging high prices to patients. This means that new commercial medicines for cancer, blood pressure or AIDS are often priced out of reach of the poor. Proposed alternatives to increase access would replace drug companies' exclusive patents, substitute company profits with public research prizes and develop a parallel, publicly funded pharmaceutical system. There has been no progress on this.

The second goal is to increase government efforts to develop medicines for non-profit neglected diseases such as malaria, sleeping sickness and parasitic infections. But this not-for-profit work was well under way long before the WHO announced its pilot projects. Governments and philanthropists were investing US$3 billion a year into research for neglected diseases, with more than 360 pharmaceutical projects in the pipeline, thousands of back-up research projects and some 40 neglected-disease drugs already being used by patients in the developing world. The pipeline was good, but the work needed more funding and better coordination between the many funders.

The WHO's motives were (and remain) good, but it made the crucial mistake of conflating these two issues. Almost from the start, the agency assumed that profits and patents were the problem not only for access to commercial medicines (they are), but also for not-for-profit public research (they aren't). And, wanting to take on the commercial model, but unwilling to front up to industry, the WHO chose instead to play out its fight in the neglected diseases. The agency insisted that the pilot projects — all publicly funded, non-profit efforts — must target both goals: to make new medicines and to trial R&D models designed to break commercial patents and profits.

This makes no sense. Innovative R&D models that replace sales revenues, remove monopoly commercial patents or replace commercial profits with public funding can be piloted only in areas that have sales revenues, patents and profits — that is, in commercial areas. Neglected diseases by definition have no profits (that is their problem in the first place), and their R&D has never been funded from sales revenues (which do not exist in impoverished regions). Neglected-disease R&D is a non-profit area that has always been funded by government and philanthropic grants, with almost all the resulting products sold in poor countries at low- or no-profit prices. The WHO has set up a false battle.

THE AGENCY'S **APPROACH** TURNS TWO POTENTIALLY SOLUBLE **PROBLEMS** INTO ONE INSOLUBLE **GORDIAN** POLICY KNOT.

The agency's approach has knock-on effects. First, it creates deep confusion among governments and public-health experts, and turns two potentially soluble problems into one insoluble Gordian policy knot. Funding of neglected diseases cannot be achieved by attacking commercial patents; and commercial access should not be gained by setting up a parallel public R&D system. The two problems need to be separated. The most recent World Health Assembly was painful proof of this, with the neglected-disease proposals sinking to the bottom along with the commercial-access millstone.

Second, WHO actions are in danger of helping to defund and shut down the successful not-for-profit pipeline of drugs for the developing world. Government funding of not-for-profit product development is down $120 million on 2009, and product-development partnerships are down $156 million. Instead of encouraging donors, the WHO is sending a message that a new R&D approach is needed, and setting up its own R&D pilot in competition.

The agency's pilot process runs a poor second to the existing neglected-disease R&D pipeline. It has no clear priorities, and the final eight projects add little value: half are re-announcements of existing work; others are low-innovation tweaks rather than priority medicines. The pilot proposals should be dropped — they are bad policy all round. The agency should instead focus on real solutions to improve access to commercial medicines — the most pressing need for the world's poor. And it should get behind the existing neglected-disease pipeline and urge funders to do the same. ■

➲ NATURE.COM
Discuss this article online at:
go.nature.com/ndrqmy

Mary Moran *is executive director of Policy Cures in Sydney, Australia.*
e-mail: mmoran@policycures.org

RESEARCH HIGHLIGHTS

Selections from the scientific literature

PROTEINS

Antifreeze protein has a heart of ice

A molecule that prevents the blood of winter flounder (*Pseudopleuronectes americanus*) from freezing is the first protein discovered to have a water-filled core.

Peter Davies of Queen's University in Kingston, Canada, and colleagues crystallized a protein called Maxi, which binds ice crystals to prevent larger ice structures from forming. Unlike most proteins, which have hydrophobic inner surfaces that exclude water, the researchers found that Maxi's core is full of water.

The inner surface of each Maxi protein holds about 400 water molecules in ordered structures layered between the protein's chains. This water structure sticks outside the protein, where it appears to bind to ice.

Science 343, **795–798 (2014)**

JOAN MONTANYA *ET AL.*/AGU

ATMOSPHERE

Turbines shoot upside-down lightning

Wind turbines emit lightning flashes upwards, producing these electrical discharges at regular intervals relative to the turbine's rotation, and can do so tens of kilometres away from an active thunderstorm area.

Joan Montanyà at the Polytechnic University of Catalonia in Terrassa, Spain, and his colleagues plotted radio emissions from lightning strikes detected by a mapping array system installed on the east coast of Spain. Later, a high-speed video was used to capture the flashes (**pictured**). The authors found that turbine blades send electrical discharges upwards in synchronization with their rotation; these discharge episodes lasted for more than an hour under certain storm conditions. The results confirm that rotating wind turbines can initiate lightning more easily than static objects, the authors say.

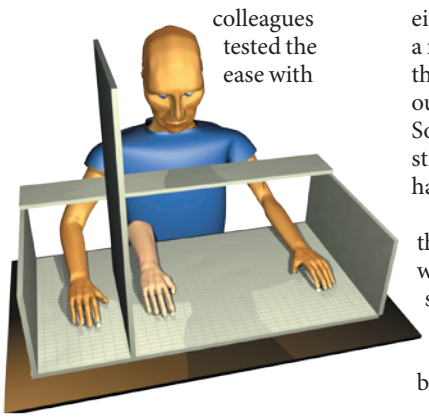
J. Geophys. Res. Atmos. **http://doi.org/rfj (2014)**

NEUROLOGY

Implications of a gentle caress

Soft, slow stroking of the skin contributes more to a sense of body ownership than other types of touch.

Haike van Stralen at Utrecht University in the Netherlands and her colleagues tested the ease with which study participants could be deceived into feeling that a rubber hand was part of their own body. Volunteers watched the fake hand being stroked quickly or slowly by either a cosmetic brush or a rough plastic cloth, while their real hand was touched out of sight (**pictured**). Soft, slower stroking gave a stronger illusion that the fake hand was their own.



RODERIK VAN HEIJST/ELSEVIER

The authors propose that the C tactile nerve fibres, which are activated by soft stroking of limbs at around 3 centimetres per second, may modulate how the brain integrates information about the body's limbs from different senses (such as sight and touch).

Cognition 131, **147–158 (2014)**

REGENERATIVE MEDICINE

Stem cells make muscles stronger

Manipulating muscle stem cells in older people could promote regeneration and prevent muscle breakdown.

Bradley Olwin at the University of Colorado Boulder and his colleagues demonstrated that a protein called p38 prevents stem cells in old muscles from renewing. When the authors took muscle stem cells from old mice and treated the cells with drugs to suppress p38, this allowed the stem cells to respond to growth signals and to replicate themselves.

In a separate study, Helen Blau and her colleagues at Stanford University in California grew old muscle stem cells on a gel while treating them with p38 inhibitors. The researchers then transplanted the stem cells into the muscles of living old mice, in which they began repairing the degenerating muscle tissue. When the group tested these muscles, they

were stronger in response to a stimulus than muscles that had received non-treated stem cells.
Nature Med. **http://doi.org/rhg; http://doi.org/rhh (2014)**

ANIMAL BEHAVIOUR

Picky spiders prefer virgins

Male black widow spiders (*Latrodectus hesperus*) prefer their female mates to be healthy and chaste — a rare demonstration of mate selection by males.

Emily MacLeod and Maydianne Andrade at the University of Toronto Scarborough, Canada, studied whether male black widows are fussy about females. Males overwhelmingly chose to mate with well-fed females who had not previously mated, both in controlled field studies and in the wild.

Male spiders' preference for well-fed virgins could be strong enough to cause females to have evolved traits, such as the production of sex pheromones, that entice males and advertise the female's status, say the authors.
Anim. Behav. 89, **163–169 (2014)**

CANCER

Unmasking the real risk genes

Cancer researchers have found part of the answer in the case of the 'missing heritability' — the mismatch between genetic disease risk and common genetic variants.

It has long been thought that common variants might be weakly linked to disease because they are co-inherited with rare, nearby genetic variants. These variants are much more predictive of disease, but are themselves too rare to be discovered, even in studies with large sample sizes.

Zsofia Kote-Jarai of the Institute for Cancer Research in Sutton, UK, and her colleagues report the first evidence for this 'synthetic association' in cancer. They show that four common genetic variants in or near the *HOXB13* gene are associated with a roughly 30% increase in cancer risk and are almost always inherited with a rare gene variant, which itself predicts an approximately 400% increase in risk for the disease.
PLoS Genet. 10, **e1004129 (2014)**

CONSERVATION

Birds should fear windows

Collisions with windows are the second-largest human-related source of US bird mortality (after house cats), with low-rise buildings and homes responsible for most deaths.

Scott Loss of the Smithsonian Conservation Biology Institute in Washington DC and his colleagues examined 23 studies of bird collisions, allowing them to estimate that between 365 million and 988 million birds are killed every year by hitting buildings.

Despite the problem of bird strikes on skyscrapers being well publicized, the team found that these impacts represented less than 1% of deaths.

Several bird species, including the golden-winged warbler (*Vermivora chrysoptera*) and the painted bunting (*Passerina ciris*) were particularly collision-prone.
Condor 116, **8–23 (2014)**

MATERIALS

A bulk graphene mimic

Physicists have identified a material that can conduct efficiently in multiple layers.

After the rush of interest in the atom-thick layers of carbon known as graphene, materials scientists turned to metal dichalcogenides — layered compounds that are also good conductors of electrical current. Rhenium disulphide now joins the promising candidates in this family.

Junqiao Wu at the Lawrence Berkeley National Laboratory in California and his colleagues discovered that bulk samples of the material have a direct bandgap, a gap in energy levels that can be used to absorb or emit light efficiently. Other members of the family have this type of bandgap only when isolated in monolayers.
Nature Commun. 5, **3252 (2014)**

COMMUNITY CHOICE

The most viewed papers in science

OBESITY

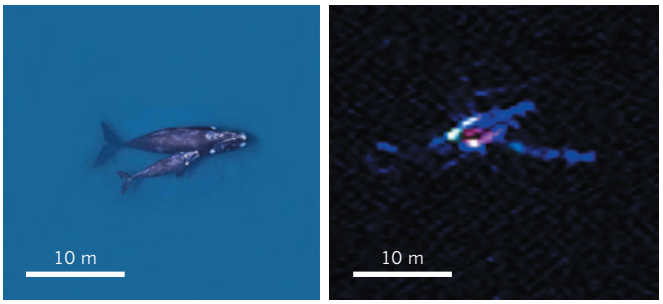
Conflicts mar studies of sweet drinks

HIGHLY READ on plosmedicine.org 13 Jan–12 Feb

Research exploring possible links between sugar-sweetened drinks and weight gain could be biased by financial conflicts of interest.

Maira Bes-Rastrollo of the University of Navarra in Pamplona, Spain, and her colleagues combed three databases for systematic reviews about the association between sugar-sweetened beverages and weight gain. The team found 18 conclusions in 17 such reviews, and 6 of those papers contained disclosures of financial ties to the food industry.

Of the 12 conclusions that had no reported financial conflict of interest, 83.3% said that consumption of sugar-sweetened beverages could be a risk factor for weight gain. But the same percentage of conclusions with a reported financial tie to industry said that there was insufficient evidence of a link.
PLoS Med. 10, **e1001578 (2014)**



DIGITALGLOBE/BAS

REMOTE SENSING

Counting whales from space

Researchers have for the first time counted whales from space, tallying 55 "probable" southern right whales and several other whale-like objects off the coast of Argentina.

Peter Fretwell and colleagues at the British Antarctic Survey in Cambridge, UK, analysed a single high-resolution WorldView2 satellite image (**pictured**, right; an aerial photograph is shown for comparison, left) from Earth-imagery company DigitalGlobe. The image covered Península Valdés, a breeding area for a major population of southern right whales (*Eubalaena australis*).

The team found that all nine available spectra could reliably identify whales at the surface, and a far-blue 'coastal band' that penetrates deeper into the water pinpointed possible whales below the surface. An automated detection system found 89% of the whales that had been classified manually as probable sightings.
PLoS ONE 9, **e88655 (2014)**

NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

RESEARCH

Stem-cell inquiry

An investigation has been launched into last month's groundbreaking reports that simply squeezing cells or bathing them in acid can reprogram them into an embryonic state. The RIKEN Center for Developmental Biology in Kobe, Japan, said on 14 February that it was looking into alleged irregularities in the work of biologist Haruko Obokata, who works at the centre and who led the studies, which were published in *Nature*. The inquiry follows some failed attempts to replicate the results and allegations about problems with images in the papers. *Nature* is also investigating. See go.nature.com/6cagqv for more.

Neutrino study

An experiment run by the Fermi National Accelerator Laboratory in Batavia, Illinois, has detected neutrinos beamed from an unprecedented distance of 800 kilometres, according to a report released on 11 February. The further that the subatomic particles travel, the more researchers can learn about them. The NuMI Off-Axis Electron Neutrino Appearance (NOvA) experiment also hopes to shed light on why the Universe has more matter than antimatter.

NUMBER CRUNCH

462,000

Amount in US dollars pledged by donors to the Immunity Project, a crowd-funded initiative to develop an HIV vaccine that has sparked debate among scientists. See go.nature.com/hwcnwu for more.

PHOTOSHOT

Politicians vow to get tough on poaching

A major political meeting in London has agreed to ramp up the fight against illegal wildlife trafficking, in the face of a huge rise in poaching. Countries including Kenya, Gabon, Tanzania, the United States, China, Germany and the United Kingdom agreed on 13 February to treat activities linked to poaching as a 'serious crime' — a technical definition meriting tough penalties for criminals — among other measures. On 11 February, the United States also announced a domestic ban on selling African elephant ivory. More than 20,000 elephants and 1,000 rhinoceroses were poached in the past year in Africa. See go.nature.com/qjupqc for more.

POLICY

GM maize

Europe may allow farmers to grow a genetically modified (GM) variety of maize (corn) after a proposal to approve the crop did not receive enough opposition to be quashed at a meeting of European Union member states on 11 February. Of 28 countries, 19 voted against the move, but their weighted contributions did not add up to a decisive majority. The European Commission is now legally required to approve the variety, which has been declared safe by the European Food Safety Authority in Parma, Italy. The crop, Pioneer 1507, produces a pesticide and would become the third GM crop to be approved in the European Union. See go.nature.com/hez8v5 for more.

Open access

The publisher of *Science* is to launch its first open-access journal in early 2015. The nonprofit American Association for the Advancement of Science (AAAS) announced the online-only journal, to be called *Science Advances*, on 12 February. Fees for publishing papers would be "within industry norms", said AAAS executive publisher and chief executive Alan Leshner. The London-based Royal Society has also announced a multidisciplinary open-access journal: on 18 February it revealed that *Royal Society Open Science* will launch in Autumn 2014 to complement the society's existing open-access content. See go.nature.com/mtlcdd for more.

EU–Swiss row

European Union–Swiss research is under strain after a Swiss vote in favour of immigration quotas led the European Commission to suspend talks on the nation's participation in Europe's €80-billion (US$110-billion) Horizon 2020 research programme. Switzerland might

XINHUA/XINHUA PRESS/CORBIS

be refused its 'associate partner' status in the programme, thus limiting scientists' ability to use European Research Council grants at Swiss institutes or to lead European Union-funded research consortia. See pages 265 and 277 for more.

Irrigation call

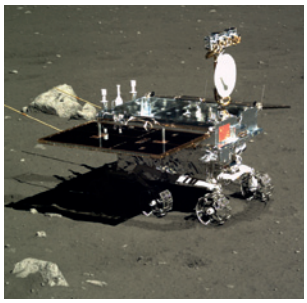
Global yields of maize (corn) could rise by 67% by 2050 if farmers in the developing world stopped tilling their soil and began irrigating their fields, says a study from the International Food Policy Research Institute in Washington DC. The report, published on 12 February (see go.nature.com/annqmt), assessed technologies that could most benefit food production in the global south. Increased funding for agricultural research also came high on the list of recommendations.

Disease control

The US Centers for Disease Control and Prevention and the Department of Defense announced on 13 February that they will team up with 26 countries, as well as agencies including the World Health Organization, over the next five years to improve global disease detection and control. They established a Global Health Security Agenda that calls for countries to increase immunizations and share data. US President Barack Obama will ask for an extra US$45 million for the programme in his budget request next month.

EVENTS

Rover resurrected



China's first Moon rover, Yutu ('Jade Rabbit', **pictured**), may yet be saved. The Chinese space agency initially said on 12 February that efforts to rouse the rover had failed after it experienced mechanical problems in late January before going into hibernation ahead of a two-week lunar night. But on 13 February, the space agency announced that it had resumed contact with Yutu. China is only the third country in the world to land on the Moon, after the United States and the former Soviet Union.

Fukushima water

Radioactive cooling water stored at the destroyed Fukushima Daiichi nuclear power plant might need to be dumped into the sea. The International Atomic Energy Agency, based in Vienna, raised the possibility of controlled discharges of pretreated water in its road map towards decommissioning the plant, which was delivered to Japan's government on 12 February.

BUSINESS

Power out

The US Energy Information Administration predicted on 14 February that the energy generated by coal-fired power plants in the United States will shrink by 20% (the equivalent of 60 gigawatts) by 2020. The reduction is the result of power-plant closures arising from competition from lower-priced natural gas and the need to modify plants to meet emission limits that take effect in April 2015. Power companies have already begun shutting down many smaller, inefficient facilities: 85 plants with a combined capacity of 10.2 gigawatts were retired in 2012.

Satnav success

Europe's fledgling satellite-navigation system, Galileo, is working well, the European Space Agency announced on 10 February. Tests of the network's first four satellites showed that the system could accurately determine positions across the planet. Over the coming year, six more spacecraft will join the network, which will eventually consist of 30 satellites. The system is planned as a rival to the US-owned Global Positioning System, and services are scheduled to start by the end of 2014.

Stem-cell patent

Woo Suk Hwang, the disgraced Korean stem-cell scientist, was granted a patent from the US Patent and Trademark Office (USPTO) for human embryonic stem cell technology on 11 February. Hwang was found guilty of embezzlement and bioethics violations in 2009 (see *Nature* **505,** 468–471; 2014). The USPTO told *Nature* that it was aware of Hwang's fraudulent past and that the terms of the patent state that his stem-cell lines must be made available on request.

COMING UP

19 FEBRUARY
NASA announces findings from its high-energy X-ray mission, the Nuclear Spectroscopic Telescope Array (NuSTAR). The observations will reveal information about supernovae.
go.nature.com/ocxk3r

22 FEBRUARY
A spacecraft weighing just 3 kilograms will hitch a ride to the International Space Station. The miniature craft, containing 100 tiny satellites, is owned by schoolchildren and space enthusiasts. The KickSat mission was funded by the crowdsourcing website Kickstarter.
go.nature.com/gzd6ni

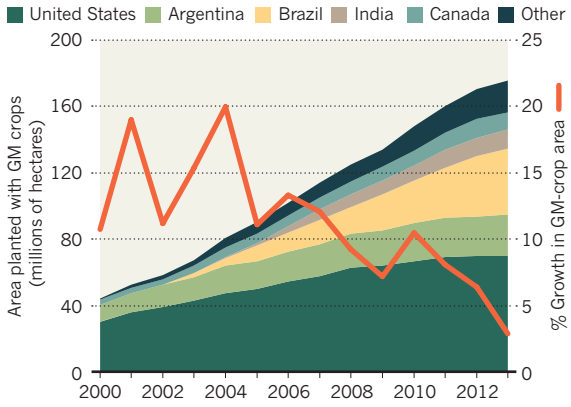
SOURCE: ISAAA

TREND WATCH

Global planting of commercial genetically modified (GM) crops rose 3% last year to 175 million hectares — the smallest-ever year-on-year percentage increase. The United States planted 70.1 million hectares, just 0.9% more than the previous year, although it did plant its first drought-tolerant maize (corn) varieties, suggesting potential for future growth. The figures were released on 13 February by the non-profit International Service for the Acquisition of Agri-Biotech Applications.

GM-CROP GROWTH SLOWS

US growth of transgenic crops may be reaching saturation point.



NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS

CHINAFOTOPRESS/GETTY

Heavy smog shrouded advertising boards in Tiananmen Square, Beijing, last month.

ENVIRONMENT

Fight against smog ramps up

Chinese government to provide incentives for heavy polluters to go green, but analysts question whether its wider air-quality strategy goes far enough.

BY JANE QIU

After decades of breakneck economic development, China is being plagued with choking pollution on an unprecedented scale. The smog over many cities reached new levels in the run-up to Chinese New Year on 31 January, creating havoc for holiday travellers. The government advised citizens to refrain from using fireworks and to stay indoors.

But change is in the air. On 12 February, China's cabinet announced that the government will implement a series of measures aimed at shifting the primary energy source from coal to natural gas and renewables; place tougher controls over emissions; and set up a 10-billion-renminbi (US$1.7-billion) fund to help companies to meet new environmental standards. A key feature will be an emphasis on using economic incentives — such as pricing mechanisms that favour cleaner alternatives to coal and crude oil, as well as taxation and requirements for investors to support only green energy companies — to encourage industry to reduce emissions and to foster the development of clean technologies, the cabinet said in a statement.

The announcement forms part of an ambitious 1.7-trillion-renminbi scheme to drastically improve air quality across China by 2017. The targets of the initiative, announced last September, include cutting atmospheric levels of PM10 — particulate matter with a diameter of 10 micrometres or less — in all major cities by 10% from the 2012 level. The government also aims to reduce the proportion of coal used in power production by nearly 2%, to improve fuel standards, to phase out highly polluting vehicles and to limit the number of cars in big cities.

Tougher targets have been set for three particularly smog-ridden regions: the Greater Beijing area, the Yangtze River Delta in eastern China and the Pearl River Delta in Guangdong province. By 2017, each must reduce its atmospheric levels of the finer PM2.5 — by 25%, 20% and 15%, respectively. "This will require a regional approach" rather than leaving individual cities to their own devices, said Zhai Qing, a vice-minister for environment, last week.

Although no details have been released about how the targets will be met, researchers say that the scheme is by far the toughest and the most ambitious in China's efforts to curb pollution. "This is the first time China has put a limit on absolute emission levels," says He Kebin, who studies air-pollution control at Tsinghua University in Beijing. Previous schemes were aimed at reducing emission intensities — the amount of emissions per unit of gross domestic product. "Although emission intensities continue to decrease, the fast economic development means that the total ▶

levels are still on the rise," he says.

To meet the targets, "China must switch from a model of economic development at all costs to one that is much more sustainable", adds He. "This is a unique opportunity to make that transition."

But the switch will not be easy, says Ma Jun, director of the non-governmental Institute of Public and Environmental Affairs in Beijing. "The resistance from industry groups and some government agencies will be fierce because of vested interest." Ma adds that institutional reforms and amendments to environmental laws will be crucial to the scheme's success.

At the moment, the environment ministry is powerless to act against many polluters, which often ignore it or happily pay fines that have little effect on their profits. The ministry "is seriously under-resourced", says Michael Walsh, founding chair of the International Council on Clean Transportation, a non-profit organization headquartered in Washington DC. The Chinese ministry has, for instance, only a few dozen staff to safeguard air quality. By comparison, the US Environmental Protection Agency has more than 1,000, Walsh says.

Analysts expect the government to reorganize ministries substantially at the annual meeting of the National People's Congress next month. Measures are likely to include slimming down powerful agencies such as the National Development and Reform Commission and the Ministry of Land and Resources, and granting the environment ministry more power and resources. Also on the agenda are amendments to the 1989 Environmental Protection Law and the 1987 Atmospheric Pollution Prevention Act. These are expected to allow the environment ministry to impose much heftier fines on heavy polluters, veto projects that fail to address environmental impacts and shut down persistent offenders.

"The resistance from industry groups and some government agencies will be fierce because of vested interest."

Political issues aside, certain scientific questions must be resolved urgently to ensure that the targets for 2017 are met, researchers say. "Air pollution in China is extremely complicated," says Zhu Tong, an atmospheric chemist at Peking University in Beijing. Smog can be caused by vehicle emissions, the burning of coal and biomass, dust and rubbish incineration. "It's unclear how their relative contribution is different in each city and varies from season to season."

As China prepares to use economic incentives and legal measures to curb pollution, it is also "making encouraging strides in promoting transparency of pollution information", says Ma. Nearly 200 cities across the country currently release air-quality information in real time, and that is set to increase to all 338 major cities next year. The environment ministry publishes a monthly list of the most polluted cities, which can affect housing prices, hit tourism and investment and provoke public outcry. Since January, the ministry has also rolled out a national monitoring system that requires more than 15,000 heavy polluters to post on the Internet their real-time emissions and discharges into rivers.

"When the information is out in the open, the polluters will be forced to act," says Ma. "The public is the strongest ally in the battle against air pollution." ■

ASTRONOMY

Missing galaxy mass found

Gravitational lensing solves puzzle from the Big Bang's echo.

BY EUGENIE SAMUEL REICH

Soon after the Big Bang, there were tiny ripples: quantum fluctuations in the density of the seething ball of hot plasma. Billions of years later, those seeds have grown into galaxy clusters — sprawling groups of hundreds or thousands of galaxies bound together by gravity.

But there seems to be a mismatch. Results released last year suggest that as much as 40% of galaxy-cluster mass is missing when compared with the amount of clustering predicted by the ripples[1]. The findings have led theorists to propose physics beyond the standard model of cosmology to make up the difference. But a reconciliation could be in the offing, using improved measurements of the cluster masses.

The mismatch was first detected by the European Space Agency's Planck spacecraft, which measured the fluctuations imprinted on the cosmic microwave background radiation left over from the Big Bang and compared them with clusters that it could see. "A lot of us are intrigued" by the discrepancy, says David Spergel, an astrophysicist at Princeton University in New Jersey, who studied the cosmic microwave background with Planck's predecessor, NASA's Wilkinson Microwave Anisotropy Probe. "There's something missing in our understanding."

Some theorists have played with the characteristics of neutrinos — ghostly, nearly massless subatomic particles — as a way of compensating. On 6 February, for example, physicist Wayne Hu of the University of Chicago in Illinois and his colleagues published[2] a theory that the mismatch could be bridged if the three known types of neutrino were significantly heavier than thought, or if there were a fourth, as yet undiscovered species of neutrino. The extra neutrino mass would have had an effect on the growth of the primordial ripples, evening them out and resulting in fewer clusters being observed today.

Now two studies, one in preparation and one posted on the arXiv preprint server on 11 February[3], suggest that clusters actually have more mass than Planck estimated — and thus

MORE ONLINE

TOP NEWS

CORBIS

Spanish flu probably came from birds, not pigs go.nature.com/xfjhzp

MORE NEWS

- First results from portable gene sequencer go public go.nature.com/qvl3bc
- PET scans trace spread of HIV analogue in macaques go.nature.com/n8iz9m
- Extant genomes yield world atlas of human admixture go.nature.com/4twiy7

NATURE PODCAST

Pathogens in bees; the Nicaragua Canal; and news from the AAAS meeting nature.com/nature/podcast

NASA/ESA/M. POSTMAN (STSCI)/CLASH

Galaxy clusters formed as a results of ripples in the very early Universe.

that there is little need for exotic physics. Both studies used gravitational lensing, a technique that weighs clusters by measuring how much their gravitational fields distort light that passes through them. "We think there's no problem," says Anja von der Linden, an astrophysicist at the Kavli Institute for Particle Astrophysics and Cosmology at Stanford University in California.

Von der Linden works on a project called Weighing the Giants, which used the Subaru telescope and the Canada–France–Hawaii telescope, both on Mauna Kea in Hawaii, to study 22 galaxy clusters also measured by Planck. It came up with an average cluster mass of 10^{15} solar masses, or about 1,000 times the mass of the Milky Way — an average that was 43% higher than Planck's estimates[3]. The other study, called the Cluster Lensing and Supernova Survey with Hubble (CLASH), used the Hubble Space Telescope to measure 25 clusters measured by Planck, and produced an estimate that was about 30% higher than Planck's.

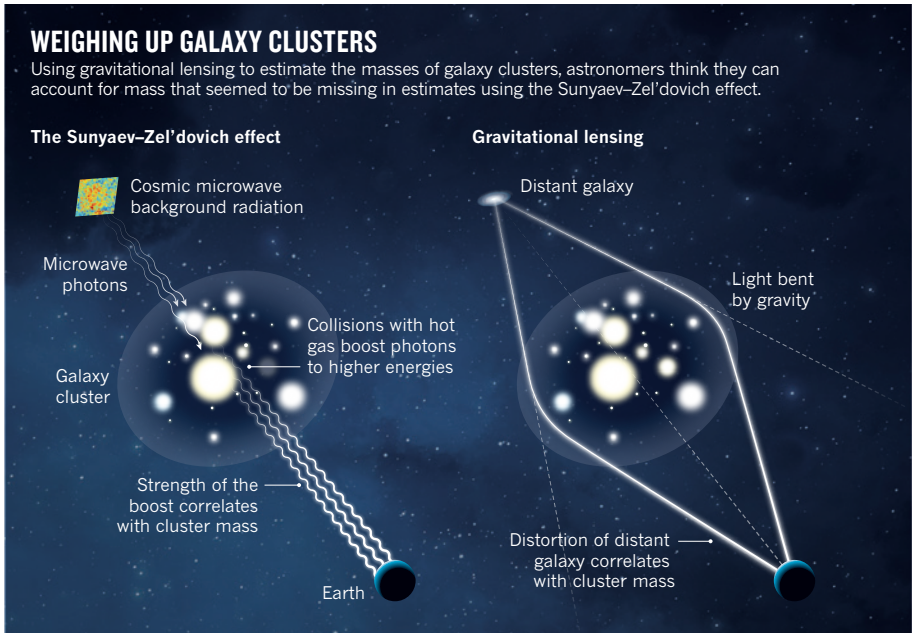
The differences seem to be attributable to the uncertain nature of Planck's estimates, which rely on a process called the Sunyaev–Zel'dovich effect (see 'Weighing up galaxy clusters'). Planck detects photons from the cosmic microwave background. On their way to the satellite, some of these microwaves pass through galaxy clusters. There, they encounter energetic electrons associated with clouds of hot gas. When the photons collide with the electrons, they are boosted to higher energies.

The strength of that signal can be correlated with the mass of all the galaxies in the cluster, because a larger cluster will trap more hot gas. But it is not a precise relationship. "That's the biggest source of uncertainty," says James Bartlett, a cosmologist at the University of Paris Diderot who is part of the Planck collaboration. He says that Planck will release an updated analysis later this year that fine-tunes its mass calibration, and may bump up the cluster masses.

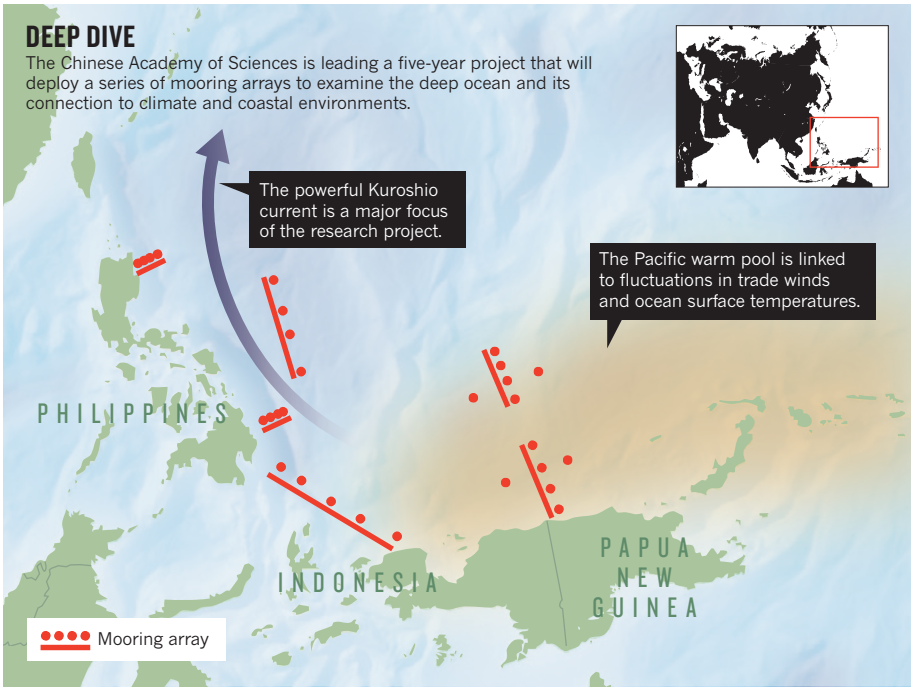
"There's something missing in our understanding."

Many astrophysicists think that the remaining discrepancies will be resolved by data from other lensing surveys that are now starting up. The US$50-million Dark Energy Survey, an optical-survey telescope at Cerro Tololo in Chile, completed its first three months of observations on 9 February. It measured hundreds of clusters and its first science results are expected later this year. Next month, a $50-million Japanese instrument, Hyper Suprime-Cam, will be used to start a large lensing survey on the Subaru telescope.

Bradford Benson, an astrophysicist at the Fermi National Accelerator Laboratory in Batavia, Illinois, says that even if turns out that no cluster mass was missing after all, it will still be important to trace the evolution of today's clusters from the original cosmic ripples to see how the effects of dark energy, a mysterious repulsive force, may have changed with time. "Powerful data sets will be the next chapter in the story," he says. ■



1. Ade, P. A. R. *et al*. Preprint at http://arxiv.org/abs/1303.5080 (2013).
2. Wyman, M., Rudd, D. H., Vanderveld, R. A. & Hu, W. *Phys. Rev. Lett.* **112,** 051302 (2014).
3. von der Linden, A. *et al*. Preprint at http://arxiv.org/abs/1402.2670 (2014).



OCEANOGRAPHY

China plunges into ocean research

Ambitious initiative targets Pacific currents, regional climate and deep-sea ecology.

BY JEFF TOLLEFSON

Six centuries ago, Chinese explorer Zheng He set sail into the Pacific Ocean with hundreds of vessels, starting a series of seven expeditions that extended China's maritime influence from Indonesia to the Red Sea. China's latest foray into the Pacific will be smaller but much more advanced. It will plumb a part of the ocean that could hold important secrets about east Asia's summer monsoon and the periodic changes in ocean temperature known as El Niño and La Niña.

Set to begin in April, the five-year research project will deploy five ships, a remotely operated submersible and an array of subsurface moorings off the eastern coasts of the Philippines and Indonesia. The region is home to the western Pacific warm pool, a patch of surface water that is formed by trade winds that influence the formation of El Niño and La Niña in the eastern Pacific. The oscillation between these periods of warming and cooling affects global climate, and the pool itself might influence regional climate events such as the Asian monsoon.

The Western Pacific Ocean System (WPOS) project is the largest single investment yet in China's growing ocean-sciences programme, says Song Sun, a marine ecologist at the Chinese Academy of Science's Institute of Oceanology in Qingdao. Over the past three years, the country has invested roughly 1.2 billion yuan (US$200 million) into Pacific Ocean science. Another $165 million will be spent on equipment and basic research costs for the WPOS project, and around 1,000 people are expected to participate. "It is a dream for Chinese marine scientists to go to the deep blue," Sun says.

"It is a dream for Chinese marine scientists to go to the deep blue."

Six arrays, comprising 29 moorings, form the core of the effort (see 'Deep dive'). The arrays will monitor ocean currents at depths of between 400 and 6,000 metres, including the start of the powerful Kuroshio current, which runs northeastwards through the East China Sea. "We have had sporadic observations in different seasons, but we've never been able to get a comprehensive view of the oceanic currents in this region," says Wenju Cai, a climate modeller at the Commonwealth Scientific and Industrial Research Organisation in Aspendale, Australia.

Dunxin Hu, a colleague of Sun's and a leader of the WPOS programme, says that the data should help oceanographers to understand the movement, temperature and nutrient load of various currents that circulate through the warm pool. Scientists are particularly interested in the Kuroshio current because it plays an important part in global ocean circulation and helps to shape coastal ecology. A key part of the ecosystem research will track the flow of nutrients into China's coastal waters and assess their influence on plankton blooms and fisheries.

The data could also prove valuable for climate modellers. Trade winds from the east have been unusually strong over the past two decades (M. H. England *et al. Nature Clim. Change* http://doi.org/rdt; 2014), which has pushed greater volumes of warm water into the region and, ultimately, into the deeper ocean. The process has helped to stall the rise in global temperatures, which have remained relatively constant since 1998, but exactly what is happening in the deep ocean remains unclear. Hu says that data from the moorings could be used to trace the heat's journey through the deep ocean. They could also reveal how the warm pool influences atmospheric convection and therefore regional climate, such as the monsoon in China.

One of the biggest challenges will be linking the deep-sea data to conditions at the ocean surface and in the atmosphere. Many of the moorings will be in fishing areas, and worries about theft and vandalism mean that they will not include surface buoys. To get the surface data, scientists will need to incorporate observations from satellites and from the global Argo network of floats that provide periodic data about temperature and salinity to a depth of 2,000 metres. "That, I think, is the missing piece," says Shang-Ping Xie, a climate modeller at the Scripps Institution of Oceanography in La Jolla, California.

Researchers also plan to investigate the deep-sea geology and ecology. They will deploy sonar-imaging equipment from a ship dubbed *Kexue* (which means 'science') to map the ocean floor. After identifying seamounts and hydrothermal vents, the team will send down the remotely operated submersible *Faxian*, which is capable of diving to roughly 4,500 metres, to study the creatures that inhabit these areas, such as sponges and exotic fish.

Peter Brewer, an ocean chemist at the Monterey Bay Aquarium Research Institute in California, helped to develop the submersible. He admires the Chinese scientists' dive into ocean science. "They are at the stage where they have all of the hardware," he says. "Now they need to train some young scientists so that they can take advantage of it." ■

POLICY

EU–Swiss research on shaky ground

Vote for immigration quotas leads to suspension of talks over Horizon 2020 programme.

BY QUIRIN SCHIERMEIER

FABRICE COFFRINI/AFP/GETTY

Swiss support for a vote to set immigration quotas (right) only just topped opposition to it (left).

World-class research facilities, lucrative funding opportunities, majestic mountains and exquisite chocolates — Switzerland has long been a land of milk and honey for scientists. But the success of a ballot to stop 'mass immigration' to the small nation of eight million has worried the international scientific community and threatens to cut Switzerland's close ties with the European Union (EU).

The binding nature of the 9 February referendum, which was initiated by a right-wing party and approved by an extremely narrow majority, will force the Swiss government to set yearly quotas to limit the influx of foreigners. Policy-makers in the capital, Bern, are drawing up a new immigration system that must be in place within three years.

The move has caused an outcry in Brussels and across Europe. Although Switzerland is not part of the EU, it maintains bilateral agreements with the bloc on key policy areas, including research and education. The country has for the past decade been an associate partner in the EU's multi-year research programmes, the latest of which, Horizon 2020, was launched last month. This allows Swiss-based researchers to apply for EU grants and lead large European research partnerships.

Since 2007, 147 Swiss-based researchers have received grants from the European Research Council (ERC), the fifth-highest tally among European countries. One of the EU's most ambitious research programmes, the €1-billion (US$1.4-billion) Human Brain Project, is hosted by the Swiss Federal Institute of Technology of Lausanne (EPFL). Overall, Switzerland punches well above its weight in terms of scientific output (see 'High productivity').

But more than half of Switzerland's scientists are foreigners, and Bern has been warned that it is breaching the bilateral agreement on free movement of people that entitles Swiss and EU nationals to live and work in any of the 28 EU countries and a handful of associate nations.

Research agreements will also be affected. In normal circumstances, Switzerland's status as an associate country in the €80-billion Horizon 2020 programme — to which it has pledged more than €3.5 billion — would probably have been easily agreed. But the outcome of the referendum prompted the cancellation of a meeting between the European Commission and Swiss negotiators last week, throwing Switzerland's participation in the programme into jeopardy.

The situation has caused grave concern among Swiss-based researchers. "We are in shock," says Jérôme Grosse, a spokesman for the EPFL. "Voters just haven't realized what dire consequences the referendum might have on Swiss–EU relations, and on research and innovation in our country."

The European Commission said last week that it expects Switzerland to agree to the free movement in Europe of citizens from Croatia, the EU's newest member, before discussions over Horizon 2020 can resume. But on 16 February, the Swiss federal councillor in charge of justice, Simonetta Sommaruga, informed Croatia and the commission that Switzerland will not grant Croatians unrestricted free right to live and work in the country. In response, the commission has suspended negotiations over Switzerland's involvement in Horizon 2020.

If Switzerland is refused its previous Horizon 2020 status, Swiss-based researchers could still participate in the programme on a project-by-project basis, as scientists from non-EU or non-associate countries such as the United States already do. But scientists would not be able to use ERC grants awarded under Horizon 2020 to do research at Swiss host institutes. And scientists based in Switzerland would not be able to lead EU-funded research consortia in the future. Nicholas Antonovics, a spokesman for Máire Geoghegan-Quinn, the European research commissioner, says that the commission will not comment at this stage on what that might mean for the Human Brain Project.

A "diplomatic effort on the highest level" will be required to prevent harm to Swiss — and European — science, says Bruno Moor, head of International Cooperation at the State Secretariat for Education, Research and Innovation in Bern. "Switzerland needs Europe just as Europe needs Switzerland," he says. "We have a responsibility to move heaven and earth to get out of this cascading crisis."

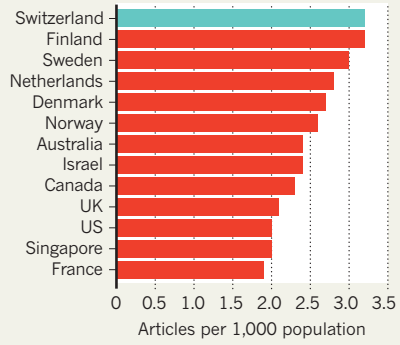
Anything short of maintaining Switzerland's status in EU research programmes would be troubling, says Daniel Höchli, director of the Swiss National Science Foundation in Bern, the main grant agency. The foundation, along with the Swiss Academies of Arts and Sciences, university rectors and the presidents of Swiss institutions of higher education, had firmly opposed the restrictions on immigration.

"Ideally, Swiss scientists would remain fully eligible to take part in Horizon 2020 for a three-year interim period," Höchli says. "This would allow us to negotiate alternative ways of mutual collaboration in case the bilateral agreement on free movement of persons, and hence research, falls."

Recruiting foreign talent to Swiss labs might get more difficult, warns Dirk Helbing, a German sociologist at the Swiss Federal Institute of Technology (ETH) in Zurich. "People will think twice about whether they want to do science in a country where foreigners feel they might not be fully welcome," he says. ■ SEE EDITORIAL P.265

HIGH PRODUCTIVITY

Between 2005 and 2009, Switzerland was a world leader in output of research papers per capita.



SOURCE: SWISS STATE SECRETARIAT FOR EDUCATION, RESEARCH AND INNOVATION

ASTROBOTIC TECHNOLOGY

The third prototype of Red Rover, the Moon explorer built by X Prize team Astrobotic.

SPACE

Moon shots stuck on Earth

Some Google Lunar X Prize contenders book launches for 2015 — but many say that is a stretch.

BY NICOLA JONES

The Google Lunar X Prize is dangling new carrots in front of the 18 private teams that are trying to put a lander on the Moon by the end of 2015. On 19 February, the prize organization announced that five teams — two US groups, one Indian, one German and one Japanese — will compete for US$6 million in 'milestone prizes'. To win the cash, prototype landers will have to demonstrate by September 2014 that they can soft-land on the Moon, move more than 500 metres once there and beam back video from the surface.

The tests will be taking place on Earth. And that is where many think that the landers will remain two years from now. Although the milestone prizes offer further financial encouragement — and catnip for the media — the goal of reaching the Moon still seems very far away. Some team members and outside observers have doubts that the $20-million main prize, which aims to stimulate the private market for getting to the Moon, will be won at all.

Jonathan McDowell, a space historian and astronomer at the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts, says that the technological hurdles are too high and the financial incentives too low. "The Google Lunar X Prize is one of the least promising things I've seen come out of the private space industry," he says. "In two years, I just don't see it from any of these teams."

The X Prize Foundation, based in Culver City, California, will probably have to extend the 2015 deadline, says Wolfgang Demisch, a retired aerospace consultant in New York who has served on the US National Science Foundation's Next Generation Launch Technology panel. But Alex Hall, senior director for the Lunar X Prize, says that the organization is not contemplating any such extension.

The rules and deadlines have already been adjusted since the prize was first announced in 2007. The original deadline of 2012 was extended to 2015 in November 2009. And in November 2013 — when it was clear that China was going to put its rover, Jade Rabbit (or Yutu), on the Moon — organizers withdrew a provision that would have reduced the prize money if a government-sponsored mission beat the X Prize teams to the punch (see go.nature.com/ftdob8).

Part of the problem has been convincing investors to put up cash. When the $10-million Ansari X Prize for achieving repeatable spaceflight with a manned craft was won in 2004 by Scaled Composites of Mojave, California, it was easy to see a financial motivation beyond the prize itself. The sponsors of the winning team founded Virgin Galactic, based in Las Cruces, New Mexico, which has been taking deposits for space-tourist tickets costing $250,000 since 2005.

To clarify the business outlook, this time, the X Prize Foundation commissioned a study by UK consultancy London Economics. The study concluded that by 2025 there would be a $1.9-billion lunar market for everything from hauling payloads to hardware development. Payload delivery is a big part of the plan for Astrobotic, a spin-out company from Carnegie Mellon University in Pittsburgh, Pennsylvania, and one of the teams competing for the milestone prizes. "Think of us as a FedEx to the Moon," says the company's chief executive, John Thornton. It already has a handful of payload contracts — for example, to deliver human ashes to the Moon for Celestis, a company that does 'memorial spaceflights'. Astrobotic plans to charge $1.2 million per kilogram for a soft-landing delivery, and less for a non-controlled drop.

But most teams are hoping for bigger customers — in particular, NASA. And the agency is indicating its interest: in 2010, it handed out six contracts (including four to X Prize teams) worth up to $30 million over five years to collect data on everything from lander test-flight results to the properties of lunar soil. Then, last month, NASA announced a programme that will offer free technical expertise, equipment and facilities to companies developing lunar landers. The only academic X Prize team, Lunar Lion of Pennsylvania State University in University Park, says that it is using the competition to gain an introduction to the world of NASA mission contracts. But McDowell points out that money from NASA "is very dependent on whims of Congress".

The X Prize competitors expect to be able to complete their first Moon mission for tens of millions of dollars. That would be cheap compared to past NASA missions, such as the Mars Pathfinder mission that put a small rover on the red planet in 1997 for $265 million. But even with the lower price tags, none of the 18 teams has reported raising sufficient funds. Only four teams — Astrobotic, Lunar Lion, Moon Express and Barcelona Moon — have booked launches for 2015 (see go.nature.com/ftdob8). And only Astrobotic has reported testing a full prototype lander. A few of these teams confessed to *Nature* that the 2015 prize deadline will be tight. "It's not impossible," says Lunar Lion team leader Michael Paul. "Feasible is a different word."

Most intend to continue with their business plans, with or without the X Prize. "We'll still be going after the Moon," says Thornton. "No matter what." ■

ACADEMIC APPOINTMENTS

Lone hunger striker spurs Nepal to action

Country's system of political patronage in science exposed.

BY SMRITI MALLAPATY

A hunger strike by a senior academic in Nepal has forced the government to backtrack on its system of political patronage in academia and highlighted the damage it has inflicted on medical services, education and research.

Govinda KC, a physician and professor at the Institute of Medicine (IOM), part of the leading Tribhuvan University (TU) in Kathmandu, began a protest fast on 8 February. His demands included the reinstatement of the medical institute's former dean, or the appointment of a new one; disciplinary action against staff involved in granting unregulated licences to private medical colleges; and the establishment of public medical colleges in rural areas. Dozens of his colleagues staged a walkout in solidarity, causing outpatient services at TU Teaching Hospital to grind to a halt.

KC's hunger strike was triggered by the resignation in December of the IOM dean, Prakash Sayami, following pressure on him to grant affiliations to private medical colleges. Such colleges must be formally affiliated to a university, and can become profitable businesses through student fees once this is achieved. Sayami had become concerned about the quality of education that they offered.

On 11 February, within hours of being sworn in, Prime Minister Sushil Koirala set a 24-hour ultimatum for TU to appoint a new dean. The university duly selected senior physician Rakesh Srivastav — a decision welcomed by KC, but one that was not enough for him to call off his protest (he had ended previous hunger strikes after being promised reforms).

"I am fasting because I want to end the corruption in the education and health institutions," he said. "I am confident that my demands will be met." KC finally ended his fast on 15 February, after the government agreed to meet almost all of his demands.

Political appointments in academia are not unusual in Nepal. "Members of the leadership team in universities are selected by the ruling party or coalition," says Kedar Bhakta Mathema, a former vice-chancellor of TU. From executive level to department heads, "it is people who can push and pull the levers of political power that can get into these positions", Pitamber Sharma, a former TU professor, told *Nature*.

But by sidelining academic merit, the research culture in Nepal has suffered, academics say. Leading positions are given to individuals who are unable to bring in research funding, argues Kamal Krishna Joshi, another former TU vice-chancellor. With government funding and student fees providing the bulk of the universities' limited funds, researchers have had to rely on their own connections and efforts to bring in cash, he says.

Political meddling has also exacerbated an uncomfortable reality in the country. "Research was never taken seriously in academic institutions in the past, it is not taken seriously even now," says Sharma. Worst affected are young scientists, many of whom have left to study and work abroad.

"I am fasting because I want to end the corruption in the education and health institutions."

Most of the funding for the Nepal Agricultural Research Council and the Nepal Academy of Science and Technology, both in Kathmandu, comes from the government and is spent on salaries. A 2010 survey by the academy found that 68% of participants claimed to have insufficient equipment for their research needs. Sharma adds that, from his own observations, Nepal is at a substandard level on many measures, including the quality of its journals and the number of publications in international journals.

But there are some positives. Last year saw spending on agricultural research — a crucial sector in Nepal — more than triple compared with 2009, from 520 million to 1.75 billion Nepalese rupees (US$5.25 million to $17.6 million).

But how much KC's fast will change things remains to be seen. Hira Bahadur Maharjan, the current TU vice-chancellor, denies claims of party interference. "In a democracy, everyone carries a political ideology. Appointed officials just happen to have a certain ideology, which is not related to their appointment," he says. The government refused to comment on the issue. ■

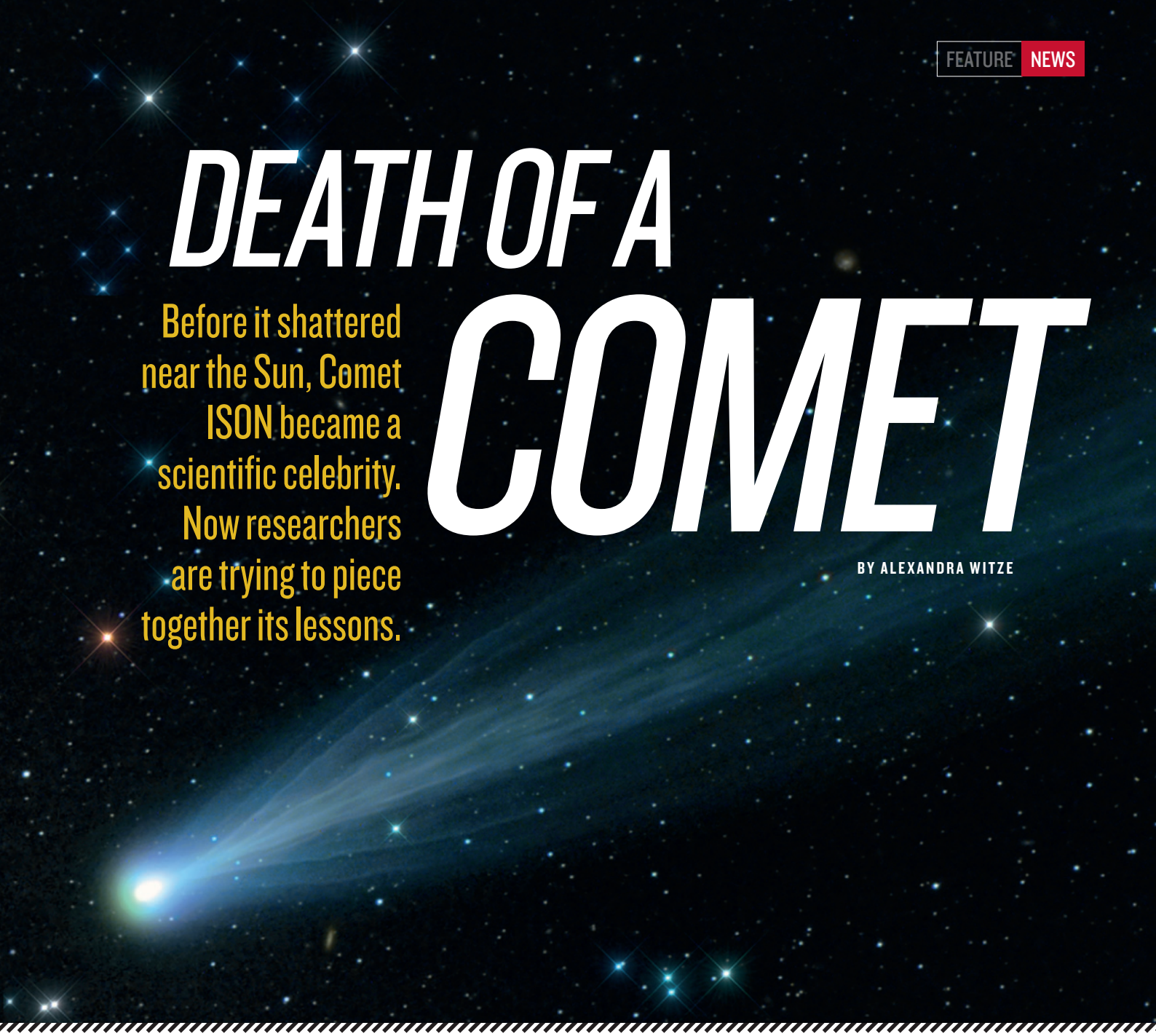
CORRECTION

The News story 'Iconic island study on its last legs' (*Nature* **506,** 140–141; 2014) wrongly gave the name of Michigan Technological University as Michigan Technical University.

DEATH OF A COMET

Before it shattered near the Sun, Comet ISON became a scientific celebrity. Now researchers are trying to piece together its lessons.

BY ALEXANDRA WITZE



DAMIAN PEACH

Near the banks of the Potomac River, in an office cluttered with craft-beer coasters and a *Doctor Who* mug, Karl Battams keeps watch for daredevil comets that skim just above the surface of the Sun.

A decade ago, when the astrophysicist joined the US Naval Research Laboratory in Washington DC, he had no deep interest in comets. But he was pressed into service because the lab operates instruments on two solar-physics missions that can spot objects passing very close to the Sun. They have detected some 2,600 such 'sun-grazing' comets so far, and it is part of Battams' job to catalogue those discoveries. He is the only dedicated sun-grazing-comet tracker in the world. "Hopefully, I'll be getting a summer student," he says one overcast January morning. "But pretty much it's just me."

All of the Solar System's comets travel around the Sun, but sun-grazers are those that fly within about three solar radii of the star's centre (some 1.4 million kilometres above its surface). Battams rose to fame last autumn as the public face of a research group tracking the most famous sun-grazer of all, Comet ISON. As ISON sailed into the inner Solar System, expectations grew quickly among astronomers and amateur skywatchers. Many hoped that it might survive its close passage to become a dramatic sight in the night sky — and continued fodder for scientific study. Instead, the comet disintegrated spectacularly in November, just hours before it was set to sweep past the Sun.

Scientists are left wondering why ISON suffered the fate it did. Early results suggest that it may have been just too small and too volatile to survive the Sun's searing heat (M. M. Knight and K. Battams Preprint at http://arxiv.org/abs/1401.7028; 2014). ISON was a tiny, gassy comet making its first ever trip to the inner Solar System — a combination that may have doomed it from the beginning.

Yet its death could mark a renaissance for the study of sun-grazing comets. ISON was spotted quite far out in the Solar System, and its unusual trajectory allowed spacecraft orbiting Earth, Mars and Mercury to photograph it from many vantage points. That made ISON the most studied sun-grazer

yet. What researchers have learned so far suggests that sun-grazers have a lot to reveal about the diversity of comets, and how hard it is to predict what they might do. Even as they wring findings out of the ISON event, astronomers are gearing up for the next close cometary encounter, later this year.

The sheer amount of observational firepower involved in studying ISON set a new standard for coordinating a flotilla of spacecraft and ground-based telescopes. "It was about bringing all of it together," says Battams. "That's never been done before."

NEVER BEFORE HAVE ASTRONOMERS WATCHED A DYNAMICALLY NEW COMET COME SO CLOSE TO THE SUN.

BLAZE OF GLORY

For centuries, skywatchers have recognized objects that disappear into the Sun and re-emerge on the other side. In 1687, Isaac Newton published the first calculations of a sun-grazer's orbit, showing that the great comet of 1680 moved according to his laws of gravitation. But it was not until the era of satellites that people could watch sun-grazers up close.

Amateur astronomers discover most sun-grazers, just days before they pass through the Sun's atmosphere, by trawling through images taken by the Solar and Heliospheric Observatory (SOHO) spacecraft. Launched in 1995, SOHO stares at the Sun with a set of three US Navy-built coronagraphs that block out the central disk of the Sun, allowing astronomers to see details in and around its blazing outer atmosphere. Once they have found a candidate comet, the amateurs alert Battams.

Most sun-grazers belong to the 'Kreutz' family of comets, named after Heinrich Kreutz, a nineteenth-century astronomer who calculated many of their orbits. Kreutz comets probably trace their origins back to a single comet that broke apart millennia ago. On each pass near the Sun, the fragments either squeak past it and survive, or plunge to a fiery doom if they come too close. Every week or so, Battams watches one of these comet pieces incinerate itself.

In 2011, however, a Kreutz comet swooped just 140,000 kilometres above the solar surface and survived — temporarily. After zipping through the Sun's upper atmosphere, Comet Lovejoy remained intact and put on a spectacular show in southern skies. Days later it fell apart, probably ripped to pieces by tidal forces induced by the Sun's powerful gravitational pull. Researchers took unprecedented measurements of Lovejoy's tail fluctuating in the Sun's powerful magnetic fields during the comet's brief passage through the solar atmosphere (C. Downs *et al. Science* **340,** 1196–1199; 2013).

In September 2012, the spotlight shifted to another sun-grazer when a pair of Russian astronomers discovered a tiny dot in the sky in the constellation Cancer. Because their telescopes belonged to the International Scientific Optical Network, the comet was dubbed ISON.

At the time, ISON was nearly one billion kilometres from the Sun, out past Jupiter. That is much farther than most comet discoveries, because comets become more active and more visible the closer they get to the Sun. The early discovery hinted that ISON was either monstrous in size or surprisingly active, prompting expectations that it would become a historic sight in the skies as it reached the inner Solar System.

Doubts emerged in April 2013, when the Hubble Space Telescope photographed ISON. Although the observing team could not measure the comet's nucleus directly, the researchers concluded that the amount of water spraying off the icy core suggested that it could be no more than 6 kilometres across, a little less than average.

Another opportunity to measure the comet came in late September as it flew past Mars. The High Resolution Imaging Science Experiment (HiRISE) on the Mars Reconnaissance Orbiter (MRO) — the sharpest-eyed camera ever to fly beyond Earth orbit — swung to look at ISON, and took a set of grainy, black-and-white photographs. These turned out to yield a better size estimate: ISON could be no more than 1.2 kilometres across. "It was a little thing," says Carey Lisse, who studies comets at the Johns Hopkins University Applied Physics Laboratory in Laurel, Maryland.

That worried astronomers. The smaller the comet nucleus, the less likely it is to survive a close pass by the Sun. Even so, some scientists thought ISON would make it (see *Nature* http://doi.org/rdm; 2013).

ISON's fate started to become apparent in the days before it reached the Sun. By 20 November, it was spraying off massive amounts of water vapour, at rates that suggested that it was no more than 500 metres across. A few days later it entered the field of view of the coronagraphs on the twin satellites of the Solar Terrestrial Relations Observatory (STEREO), on the opposite side of the Sun from Earth. The comet continued to brighten, then faded a bit, then ominously brightened again as if it were already flaming out.

The closer it got to the Sun, the worse ISON looked. "Where we really went 'uh-oh' was the morning of close approach," says Matthew Knight, an astronomer at Lowell Observatory in Flagstaff, Arizona. On 28 November, the US Thanksgiving holiday, Knight, Battams and Lisse gathered on Kitt Peak in Arizona to watch for the comet. Knight was hoping to use a solar telescope there to take spectra of ISON as it flew 1.2 million kilometres past the Sun. He got up extra early, crunched the numbers on the comet's brightness, and realized that, overnight, it had faded beyond hope. It had shrunk to an unsustainable size — perhaps just 50 metres across — and was on the verge of shattering completely.

For hours the team, plus comet fans around the world, watched as ISON flew through the SOHO and STEREO fields of view. It grew fainter as it approached the Sun, and could not be seen at all by a third solar probe peering closer to the star. On the far side of the Sun, where the nucleus would have emerged had ISON remained intact, a ghostly cloud of remnant dust appeared and quickly faded from view.

Battams took the loss of ISON personally. "It was kind of a process of heartbreak, really," he says. "Born 4.5 Billion BC, Fragmented Nov 28, 2013", he wrote in an obituary on an ISON-observing blog. "Survived by approximately several trillion siblings, Comet ISON leaves behind an unprecedented legacy for astronomers, and the eternal gratitude of an enthralled global audience."

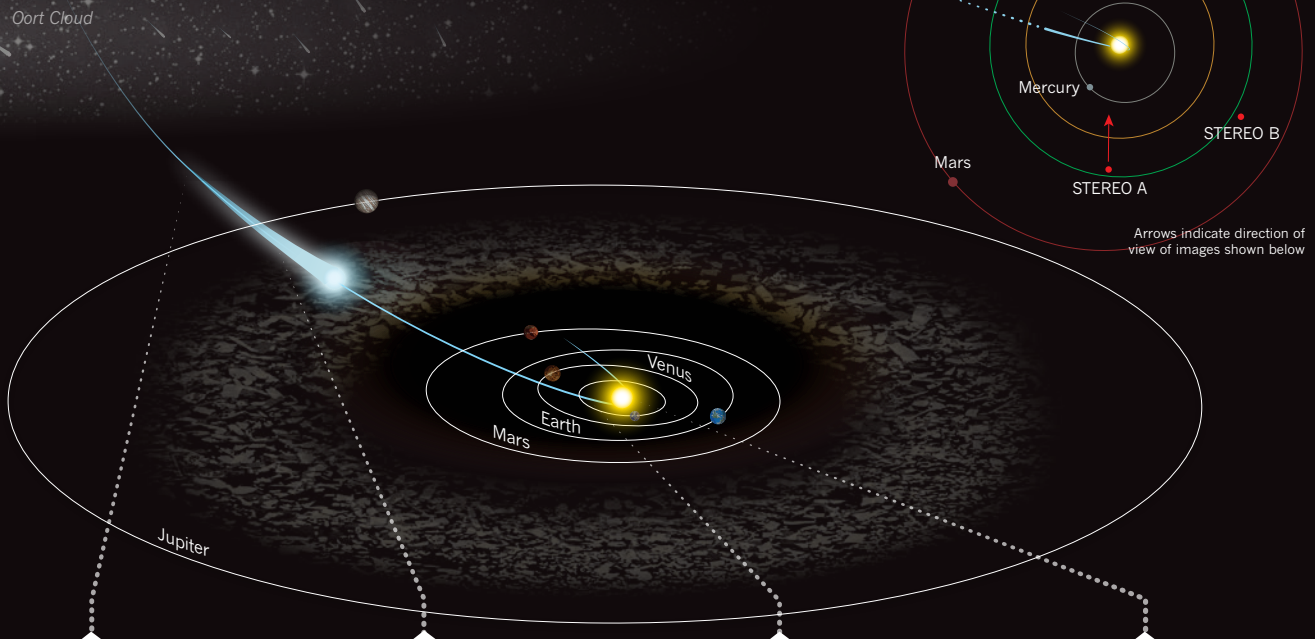
LONG TAIL

Although disappointed, Battams and other astronomers quickly moved on to glean clues from ISON's death. One major puzzle has been why ISON disintegrated and Lovejoy survived, when Lovejoy passed much closer to the Sun. The answer may lie in their ancient histories, says Knight.

As a Kreutz-family comet, Lovejoy had travelled through the inner Solar System several times before, burning off its most volatile components. But ISON was a 'dynamically new' comet that had never before visited this region of space. It hailed directly from the Oort cloud, the icy reservoir of comets beyond the orbit of Pluto. It spent most of its life in this cloud until perhaps a few million years ago, when the gravity of a passing star nudged it into a new orbit (see 'Final journey'). As it approached the Sun for the first time, volatile molecules began flying off its icy nucleus. Simple organic compounds, such as methane and carbon monoxide, would have burned off first, followed by more complex ones — much like a steak left on a grill too long. All that

FINAL JOURNEY

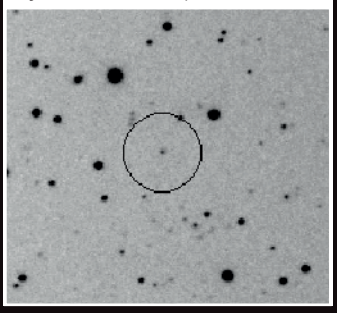
After billions of years hanging out in the distant Oort Cloud, Comet ISON plunged into the inner Solar System last year. Astronomers first caught sight of the comet in 2012, and then used an unprecedented number of telescopes in space to follow its path. Despite hopes that the comet would survive its close encounter with the Sun, ISON disintegrated in late November 2013.



SOLAR TELESCOPES CAPTURE ISON'S DEMISE

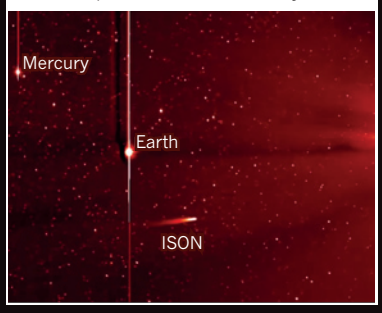
Arrows indicate direction of view of images shown below

21 SEPTEMBER 2012
Amateur astronomers discover ISON beyond the orbit of Jupiter.



10 APRIL 2013
Hubble images suggest that ISON is not as big as previously thought.

25 NOVEMBER 2013
The STEREO A satellite captures an image of ISON as it passes Earth and Mercury.



28–29 NOVEMBER 2013
A composite of images from the SOHO mission show the remnants of ISON.

SATELLITE LOCATIONS: NASA/STEREO. PHOTOGRAPHS, LEFT TO RIGHT: VITALI NEVSKI/ARTYOM NOVICHONOK; NASA/ESA/J.-Y. LI (PLAN. SCI. INST.)/ HUBBLE COMET ISON IMAGING SCI. TEAM; NASA/STEREO; ESA/NASA/SOHO

early activity explains why ISON looked so bright early on.

Never before have astronomers watched a dynamically new comet come so close to the Sun. The sequence and rate at which molecules burned off ISON are key data points for understanding the next one that comes in. "We've learned that dynamically new comets evolve and change," says Lisse. That is much more than astronomers managed to learn in 2011 from the sun-grazing Comet Elenin, which was also a relatively fresh visitor to the inner Solar System, but was too dim to provide much science before it plunged to its death.

More broadly, the ISON experience may act as a template for future comet-observing campaigns. A record 14 missions photographed the comet from space, says Lisse, who coordinated the campaign for NASA. They included the Mercury Surface, Space Environment, Geochemistry, and Ranging (MESSENGER) spacecraft and the MRO at Mars.

That experience will inform a big event this October, when a comet named Siding Spring is due to pass within 140,000 kilometres of Mars. It will be the closest such pass ever seen, and the red planet will be physically enveloped in the comet's coma, the shroud of ice and dust around the nucleus. In a manoeuvre practised with ISON, craft such as the MRO will swivel to photograph Siding Spring. "It's like a cheap comet flyby," says Lisse.

More expensively, the European Space Agency is sending its Rosetta spacecraft to land on and study Comet 67P/Churyumov-Gerasimenko in November (see *Nature* **505,** 269–270; 2014). With the combination of Siding Spring, what is popularly called CG and the aftermath of ISON, some have labelled 2014 the year of the comet.

That is just fine with Battams, who keeps a sharp eye on whatever might enter his Sun-centric field of view. "Now it's just a case of sitting and waiting," he says, "for the next interesting thing to come along." ■

NATURE.COM
For videos of ISON's journey, see:
go.nature.com/nwyxzr

Alexandra Witze *is a correspondent for* Nature *in Boulder, Colorado. She lost her Thanksgiving Day watching ISON die, and is still a little bitter.*

The US Supreme Court years ago ruled against applying the death penalty to people unable to understand the legal process. Now it must grapple with the science of how intellectual disability is measured.

SMART ENOUGH TO DIE?

BY SARA REARDON

Many psychologists say Freddie Lee Hall (circled) should not face the death penalty.

JIM GOFF/*THE TAMPA BAY TIMES*

Freddie Lee Hall loved to gamble, although he usually lost. Winning was better: then he gladly gave the money back to the friends he'd won it from, along with all the wages he earned picking fruit in rural Florida. His friends praised him for this. It made him feel good.

And Hall needed to feel good — as court documents make abundantly clear. As a child growing up in the impoverished town of Webster, Florida, he had struggled to keep up with 16 brothers and sisters, who were much smarter than he was. If he failed to understand something, his mother beat him, once while he was tied up in a bag strung over a fire. He stuttered, never learned to read and feared the dark. He was unable to live alone. "Even though he was full grown, mentally he was a child," his sister Diana told the court. "I had hoped to protect Freddie Lee from the outside world."

But the outside world found him. In 1978, Hall and his friend Mack Ruffin decided to rob a convenience store. They needed a car, so they forced 21-year-old Karol Hurst, who was pregnant, to drive into the woods, where they raped and killed her. Later, one of the pair also shot and killed a sheriff's deputy. When the two men were caught, tried and convicted of murder, the court decided that Hall was the likely ringleader. Ruffin was eventually sentenced to life in prison; Hall was sentenced to death.

Next month, after 35 years of failed appeals to have that death sentence commuted to life imprisonment, Hall will have his case heard before the US Supreme Court. His guilt is not in question: the issue is Florida's use of IQ test

scores in sentencing him to death.

A 2002 Supreme Court ruling already bars the execution of people with an intellectual disability. But Hall's lawyers are expected to argue that many US states assess mental ability using outdated measures that take little or no account of current scientific research on the subject. Florida, in particular, is one of ten states in which anyone with an IQ score of above a certain number, usually 70, is automatically considered to be intellectually competent and is therefore eligible for the death penalty. Psychologists contend that IQ tests are not precise enough to draw such a 'bright line'. Hall's IQ scores range from 60 to 80, and many states would not consider him for the death penalty, say several specialists who have reviewed his case. Documents prepared for one of his trials quote clinicians saying that Hall "is mentally retarded, has always been mentally retarded, and will be mentally retarded for the remainder of his life".

There is a great deal resting on how the Supreme Court decides this case. According to one estimate, as many as 20% of the more than 3,100 people on death row in the United States may have some level of intellectual disability (R. Coyne and L. Entzeroth *Geo. J. Fighting Pov.* **3,** 40; 1996). So a decision in Hall's favour could lead to hundreds of appeals, says Nancy Haydt, an attorney in Santa Barbara, California, who is compiling a database of pleas relating to intellectual disabilities in death-penalty cases.

But many mental-health specialists hope that the court will rule more broadly. In briefs filed in the case in December, professional organizations, including the American Psychological Association (APA) and the American Association on Intellectual and Developmental Disabilities (AAIDD), advocated for the court to set a new legal standard that reflects current research on intelligence. IQ tests were never designed to assess the criminal mind, psychologists argue. They say that the modern definition of intelligence — which includes the ability to learn and solve problems, relate to other people and function in society — is much more relevant.

CRUEL AND UNUSUAL PUNISHMENT

The confusion over the legal role of IQ began in 2002, when a Virginia man, Daryl Atkins, went before the Supreme Court to appeal his death sentence. Atkins and an accomplice had been arrested in 1996 for murdering a man. There were no witnesses, and Atkins's accomplice made a deal with prosecutors: in exchange for a life sentence, he testified that Atkins had held the gun. Atkins, who had an IQ of 59, smiled and doodled through his trial.

The Supreme Court ruled that it was "cruel and unusual" to execute a person who could not understand the consequences of his actions or the legal proceedings. They changed Atkins's sentence to life imprisonment and ruled that death is not a suitable punishment for anyone who has been diagnosed as "mentally retarded" by the standards of the American Association on Mental Retardation (now the AAIDD). This organization's definition has three criteria: an IQ score two standard deviations below average (about 70); difficulty adapting and functioning in society; and evidence that the disability began before age 18.

However, the court left it up to individual states to decide how to implement the Atkins criteria. And the paths many have chosen have raised issues with all three (see 'Bright lines').

> "LOOKING AT INTELLIGENCE AS A SINGLE NUMBER IS SUCH AN OUTMODED CONCEPT, WITH NO SCIENTIFIC VALIDITY."

Looking at the IQ standard, for example, "the states took it as 'we can define intellectual disability however we want'," says Harry Simon, an assistant federal defender in Sacramento, California. One result was Florida's bright line: an IQ score of above 70 would be enough to end a defendant's plea. Some states are even stricter: in Oklahoma, a single score above 75 from any test taken in a defendant's life automatically qualifies him or her for capital punishment.

Kent Scheidegger, legal director of the Criminal Justice Legal Foundation, a non-profit organization in Sacramento that supports the death penalty, says that the problem is the Atkins decision itself. "It created a rule of law that we have to crisply divide people in two categories — retarded and not retarded — and treat them differently," he says. "Because there isn't a clear line in reality, that is inherently problematic."

Each state now has to find some way to comply with that, Scheidegger says. The bright line rule, if nothing else, applies uniformly to all defendants, avoiding a situation whereby the outcome depends on which side has the better expert psychologist. The Florida Attorney General's office declined to comment for this article, citing the pending Supreme Court case, but its brief to the court argues that a ruling in Hall's favour would ensure that "states are constitutionally bound to vague, constantly evolving — and sometimes contradictory — diagnostic criteria established by organizations committed to expanding Atkins's reach".

Yet psychologist Keith Widaman of the University of California, Davis, says that clinicians generally oppose a bright-line test, not least because IQ tests have an error margin of roughly ten points (see go.nature.com/vmir87). Besides, most defendants have taken several IQ tests and achieved a range of scores, which can vary widely depending on the type of test and the version used. Widaman points out that one of the most commonly used IQ tests, the Wechsler Adult Intelligence Scale (WAIS), has only a few questions aimed at assessing people at the lower boundary of the normal intelligence range: right around 70.

What is more, the tests themselves have evolved over time. Conventionally, they focused on 'crystallized' intelligence, which includes factors such as a person's knowledge and ability to comprehend, say, a text. But, especially during the past decade, Widaman says, test designers have been putting an increased emphasis on 'fluid' intelligence: how well an individual can absorb new information, make judgements and reason through a complex problem.

Interpreting test scores is complicated even further by a phenomenon known as the Flynn effect: the average IQ score on a given test rises by roughly three points per decade across a population. No one is sure why; proposed explanations range from better nutrition and prenatal care to increased standardized testing in schools.

Every ten years or so, psychologists must therefore renormalize IQ tests such as WAIS so that the population's average IQ remains at 100. This means that a death-row prisoner such as Kevin Green, who in 1991 scored 71 on an IQ test last normalized in 1972, might have scored only 65 on a test normalized to the year he took it. After Green was convicted and sentenced to death in 2000, his lawyers appealed, arguing that the court should correct for the Flynn effect. Nevertheless, Green's score exceeded Virginia's bright line of 70, and he was put to death in 2008.

Whereas *Hall* v. *Florida* has focused on the standard error of measurement in IQ tests, some psychologists would like laws to reflect a broader understanding of intelligence. "Looking at it as a single number is such an outmoded concept, with no scientific validity," says Stephen Greenspan, a forensic psychologist in Littleton, Colorado, who consults on Atkins cases.

NO LIMITS

This is why the latest version of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*), used by almost all US hospitals and health providers, deliberately avoided setting any IQ number as a limit for diagnosing intellectual disability, says James Harris, a psychiatrist at the Johns Hopkins University in Baltimore, Maryland, and the lead author of the *DSM-5*'s chapter on the subject. IQ tests don't have much to say about a defendant's ability to function in society, he says. "They're not looking at what happens when someone says, 'I'm going to give you a reward if you go with me to rob the bank; it'll be a lot of fun and I'll even let you hold the gun.'" But, he says, that kind of situation is likely to come up.

This is the realm of the second Atkins criterion, 'adaptive functioning', which is given equal weight to IQ scores by the *DSM-5*. It involves factors ranging from empathy and social skills to impulse control and judgement, says Harris

— none of which bear much relation to standard IQ scores, particularly when brain damage is involved. Take the case of Michael Zack, who has been on death row in Florida since 1997 for murdering a woman he met at a bar. Over the course of several trials, his lawyers and a psychologist argued that Zack had brain damage caused by his mother's heavy drinking during pregnancy. Although fetal alcohol spectrum disorder does not necessarily lower standard IQ scores, it severely damages the midbrain, which is involved in the ability to learn from experience and to anticipate the consequences of one's actions. This could explain why Zack scored 79 on an IQ test yet was diagnosed as having the emotional maturity of a ten-year-old.

Psychologists assess adaptive behaviour using standardized tests that ask questions about real-life function, such as whether a person can tie their shoe-laces or write a cheque. These questionnaires are typically given to family members and close acquaintances rather than the individuals themselves, because people tend to exaggerate their own abilities.

But the tests are frequently misused in legal settings, says William Hennis, an attorney with Florida's Commission on Capital Cases in Fort Lauderdale. They are often given to prison guards, who say that the defendant is getting along perfectly well in maximum security. "In a situation where you get all your meals provided and are under surveillance and your entire life is organized to the minute, that's not an environment where you can do an adaptive functioning [assessment] that makes sense," Hennis says. To get a fair determination, he maintains, the test needs to be given to people who knew the defendant before he or she was in prison.

Some states have developed their own ways of measuring mental eligibility. Texas, which executes more prisoners than any other state, uses 'Briseño factors': a set of seven behavioural criteria formulated by the Texas Court of Criminal Appeals in 2004. The factors include whether family and acquaintances think that the defendant is mentally disabled, and his or her ability to answer direct questions, lie and plan ahead. The judges in the case wrote that they were following AAIDD guidelines, and described the seven factors as an attempt to be specific about "adaptive behaviour criteria [that] are exceedingly subjective". But the factors have become a lightning rod for critics. "They're criteria invented out of thin air by judges that have no validation in science," says Simon.

Such non-professional assessments can miss an important reality, he says: many disabled defendants hide behind a 'cloak of competence' by copying the actions of others or collecting books they are unable to read in order to appear literate. "A lot of people think they know what mental retardation looks like, but it doesn't look like anything," says Marc Tassé, a psychologist at Ohio State University in Columbus. Later this year, the AAIDD will release the first adaptive-behaviour test specifically designed to diagnose mild intellectual disability in young people, called the Diagnostic Adaptive Behavior Scale (DABS). "These death-penalty cases made us realize the importance of designing a test around the cut-off," says Tassé, who is heading the project.

One of DABS's novel contributions will be questions about gullibility, which is a hallmark of intellectual disability. Criminals with an intellectual disability often have an accomplice, who might have led them into the crime, says Tassé.

ENVIRONMENTAL DAMAGE

Often, the most difficult of the Atkins criteria for defence lawyers to prove is the third: has a defendant been intellectually disabled since before the age of 18? Clues can sometimes be found by looking at a defendant's early environment. Childhood neglect and abuse, for instance, can lower IQ substantially. Court documents such as Hall's brim with tales of abuse by parents and others, damaging the children's brains with blows to the head and creating traumatic memories. But records are often fragmentary or missing, forcing defence lawyers to rely on families' and teachers' subjective memories. Hennis once tried to find childhood records on a client, Dean Kilgore, who had grown up as the son of poor black sharecroppers in 1950s Mississippi. Only after days of searching did Hennis turn up 50-year-old juvenile-conviction records revealing that Kilgore had been described as "brain damaged" by others at his work camp.

Psychologists hoping for better courtroom science are encouraged by the Supreme Court's decision to hear *Hall* v. *Florida*. Conceivably, the court could rule that states should abide by the *DSM-5*'s diagnostic criteria, although Greenspan admits that such a broad ruling is unlikely. It is more probable that the court will decide that states must account for standard errors in IQ scores, or emphasize diagnoses by clinical psychologists.

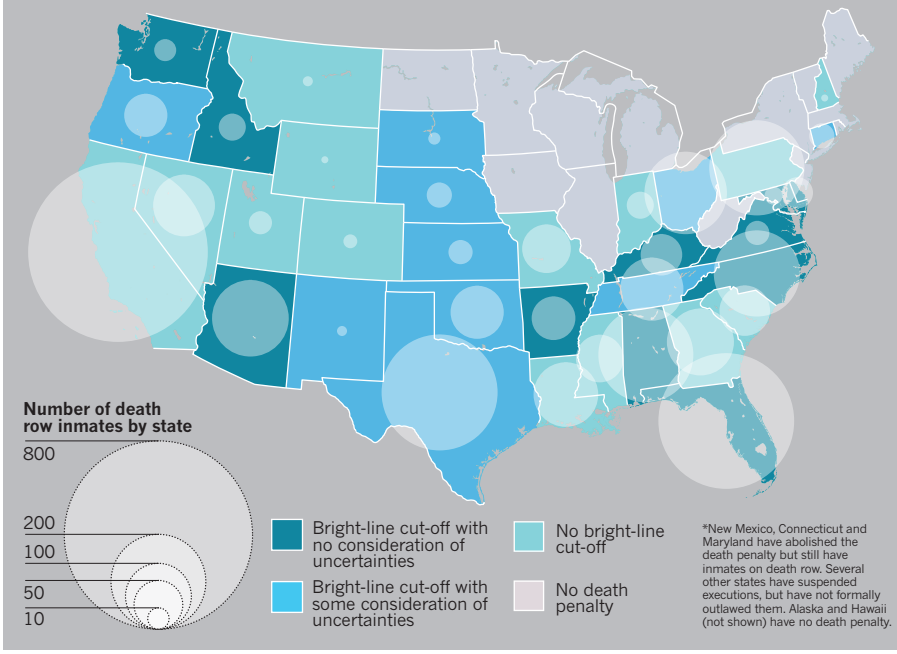
Florida contends that its method for assessing intellectual abilities meets the standards set out in Atkins and that a ruling in Hall's favour would unleash an unwarranted flood of appeals. As the state Attorney General's office wrote in its brief to the Supreme Court, "future litigation would be endless".

But advocates for some death-row prisoners are confident. "The smart better's money is that it looks good for Hall," says Lee Kovarsky, an attorney at the University of Maryland in Baltimore, who represented Marvin Wilson, a convicted killer with an IQ of 61 who was executed in Texas in 2012. In recent cases involving science, the Supreme Court has been very open to expert opinions — and there are plenty to choose from here. Among them are the briefs filed by the AAIDD and APA. Even a group of dozens of former judges and law-enforcement officials filed a brief in Hall's favour, encouraging the court to account for standard error of measurement in IQ assessments. "I think it's a very powerful statement about the dilemma that we're in," Harris says of the briefs. "We cannot reduce the life of a human being to a single number." ■

Sara Reardon *is a reporter for* Nature *in Washington DC.*

BRIGHT LINES

Every US state must decide for itself whether a convicted criminal is too intellectually disabled to receive the death penalty. Many states — often those with many prisoners on death row* — do this by drawing a 'bright line', whereby anyone with an IQ of above, say, 70 is automatically eligible for execution. In the process, some states ignore the uncertainties and limitations of IQ tests.



COMMENT

AXEL MEYER

Lake Nicaragua is the largest drinking-water reservoir in Central America and is home to fish species key to evolutionary science.

Nicaragua Canal could wreak environmental ruin

Plans for a 300-kilometre waterway joining the Pacific and Atlantic oceans need independent environmental assessment, urge **Axel Meyer** and **Jorge A. Huete-Pérez**.

Last June, the Nicaraguan government granted a concession to a Hong Kong company to build a canal connecting the Pacific Ocean and the Atlantic Ocean, through the Caribbean Sea. The HK Nicaragua Canal Development Investment Company (operating as HKND Group) signed a 50-year lease, renewable for another 50 years. It plans to break ground in December after spending this year establishing a route and conducting feasibility studies. Included in the concession are the rights to build and operate industrial centres, airports, a rail system and oil pipelines, as well as land expropriation and the rights to natural resources found along the canal route.

The Nicaraguan government says that the US$40-billion project will boost economic growth in the country — the second-poorest nation in the Americas — from 4.5% in 2013 to 14.6% in 2016. No economic or environmental feasibility studies have yet been revealed to the public. Nicaragua has not solicited its own environmental ▶

impact assessment and will rely instead on a study commissioned by the HKND. The company has no obligation to reveal the results to the Nicaraguan public.

In our view, this canal could create an environmental disaster in Nicaragua and beyond. The excavation of hundreds of kilometres from coast to coast, traversing Lake Nicaragua, the largest drinking-water reservoir in the region, will destroy around 400,000 hectares of rainforests and wetlands.

The accompanying development could imperil surrounding ecosystems. Some 240 kilometres north of the most likely route of the canal lies the Bosawas Biosphere Reserve — 2 million hectares of tropical forest that is the last refuge of many disappearing species (see 'Nicaragua carve-up'). Less than 115 kilometres to the south is the Indio Maiz Biological Reserve, with more than 318,000 hectares of tropical dry forest. Worse still, the probable canal route cuts through the northern sector of the Cerro Silva Natural Reserve.

The project threatens multiple autonomous indigenous communities such as the Rama, Garifuna, Mayangna, Miskitu and Ulwa, and some of the most fragile, pristine and scientifically important marine, terrestrial and lacustrine ecosystems in Central America.

An international community of conservationists, scientists and sociologists needs to join the concerned citizens and researchers of Nicaragua in demanding two things: first, independent assessments of the repercussions of this mega-project; and second, that the Nicaraguan government halt the project should the assessments confirm fears that this canal will yield more losses than gains for the region's natural resources, indigenous communities and biodiversity.

AT WHAT PRICE?

Many have dreamed of a canal through Nicaragua — from the Spanish conquistadors to Napoleon III. The US industrialist Cornelius Vanderbilt, the United States and the United Kingdom all had plans for such a canal by the mid-1800s, well before the Panama Canal was completed in 1914. Engineering challenges, projected costs and, more recently, competition with the Panama Canal, have prevented such plans from being realized.

"Inhabitants of all species with ancient ties to the land will be uprooted."

The most likely route[1] of the HKND's canal is 286 kilometres long and would cut an approximately 90-kilometre swathe across Lake Nicaragua, requiring a major transformation of the lake bed and local rivers. To rival the expanded Panama Canal (slated for completion in 2015) by accommodating ships of up to 400,000 tonnes, the proposed Nicaraguan waterway will be 27.6 metres deep, and the HKND has claimed that it may be an implausible 520 metres wide. Lake Nicaragua, however, has an average depth of only 15 metres. The extensive dredging required would dump millions of tonnes of sludge either into other parts of the lake or on to nearby land. Either way, the sludge will probably end up as damaging sedimentation[2].

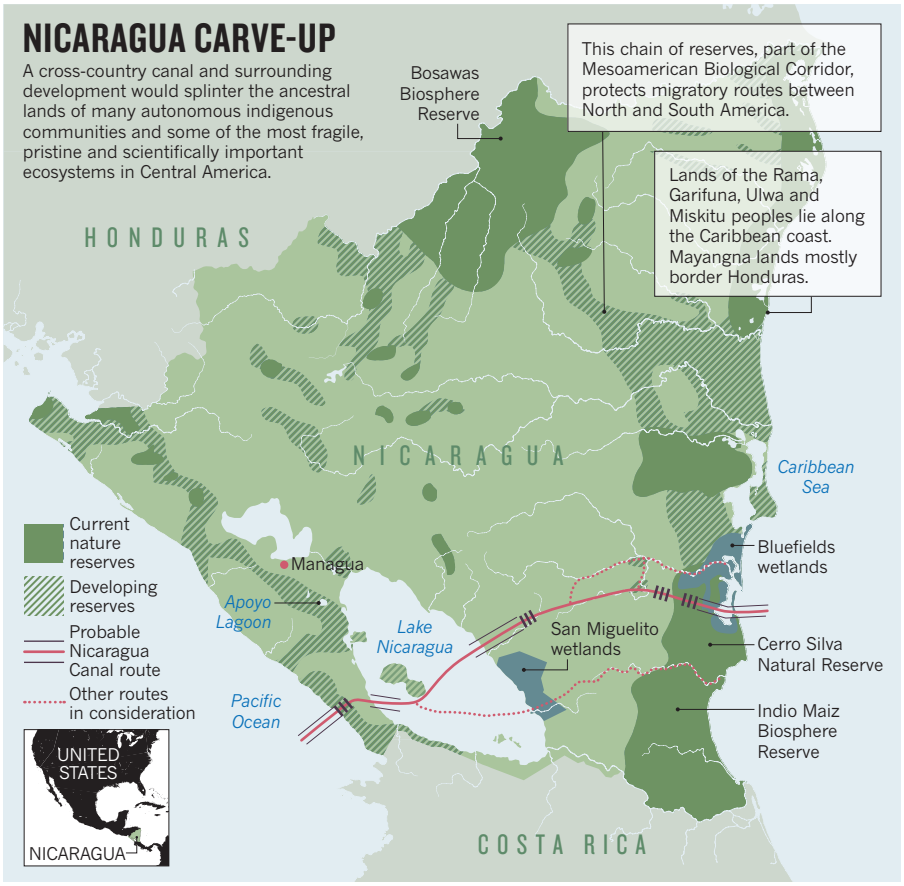
Lake Nicaragua would also serve as the reservoir for the canal's lock system, requiring dams to be constructed in an area of frequent seismic activity, which would increase the risk of local water shortages and flooding. The lake would probably suffer from salt infiltration in the lock zones, as in locks of the Panama Canal. This would transform a free-flowing freshwater ecosystem into an artificial slack-water reservoir combined with salt water. Declining populations of native aquatic fauna such as euryhaline bull sharks, sawfish and tarpon, important for sport fishing and tourism, could also suffer.

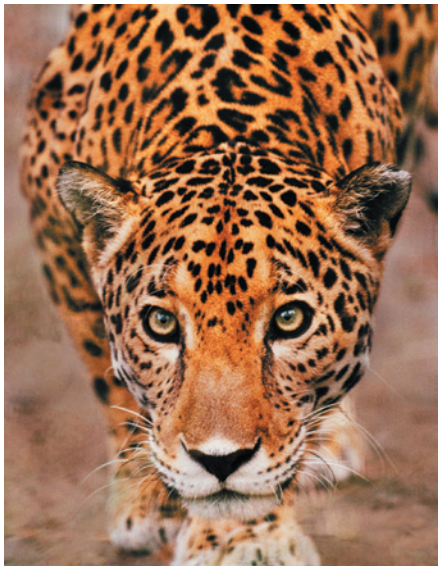
Changes in chemical composition and disruptions to dissolved oxygen levels in the water from pollutants and construction could harm numerous populations of freshwater and marine fish found nowhere else in the world[3]. Boat wakes and dredging could weaken and denude the shoreline of rivers leading inland from the new ports on both coasts[4]. This could affect the rivers Escondido, Rama and Oyate on the Atlantic side, and Las Lajas and Brito on the Pacific side.

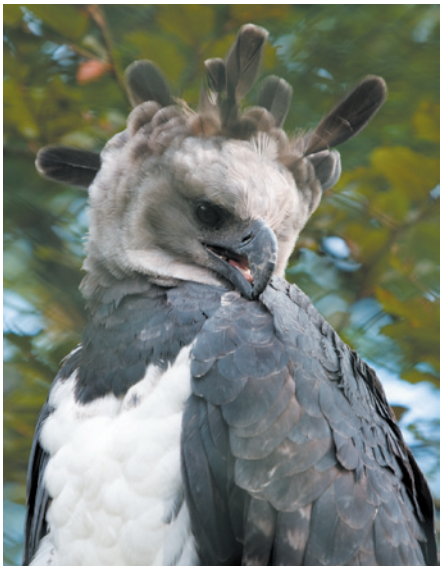
Invasive species from container bilge water are another concern. The arrival of non-native fish can have catastrophic results, as demonstrated by the dramatic decline in Lake Nicaragua's cichlid fish population since the introduction of African tilapia in the 1980s[5]. Cichlids are central to evolutionary research. Ecology and genetics studies over four decades have resulted in tens of publications involving researchers from more than a dozen countries. One study, for example, revealed that in less than 10,000 years, several species of cichlid evolved from one population in the Apoyo Lagoon, a crater lake close to Lake Nicaragua[6].

Other vulnerable ecosystems[7] in the Cerro Silva–Indio Maiz–La Selva Corridor, such as the biodiversity-rich wetlands of San Miguelito and Bluefields, will suffer from dredging, sedimentation, invasive species, emissions and other pollution. Shipping and the construction and operation of deepwater ports[8] on the Atlantic and Pacific coasts will affect the nesting and egg-laying habitats of several endangered sea turtles and threaten coral reefs and mangroves.

On land, animal populations will be artificially confined to territories carved out by the canal's infrastructure and accompanying projects, disrupting migration patterns, connectivity and ecological dynamics. Already,







CICHLIDS: AD KONINGS; JAGUAR: FRANS LANTING/FLPA; EAGLE: EDWIN GIESBERS/NATUREPL.COM

Cichlid fish, jaguars and harpy eagles are among the many species threatened by the Nicaragua Canal.

the extraordinary concentrations of endemic species in the Mesoamerican Biological Corridor are experiencing rapid habitat loss. This crucial biodiversity hotspot is a conservation system established in 1997 by Mexico and the countries of Central America to limit human activity and to create a safe migratory corridor between North and South America.

Nicaragua's Indio Maiz and Bosawas biosphere reserves — key links in this corridor — sandwich possible canal routes. Hundreds of thousands of hectares of the forests and wetlands would be cleared for the canal, destroying the habitats and food sources of already endangered species such as the Baird's tapir (*Tapirus bairdii*)[9], the spider monkey (*Ateles geoffroyi*), the harpy eagle (*Harpia harpyja*) and the jaguar (*Panthera onca*), a creature of mystical importance to Mesoamerican cultures.

The social, economic, cultural and environmental costs of population resettlement are expected to be enormous. Hundreds of villages will have to be evacuated and the indigenous inhabitants relocated. Archaeological sites along the route of the canal will be in danger too. This upheaval could reignite the civil violence that has long blighted the region. The situation is already tense as outsiders invade ancestral lands with cattle and carry out illegal logging.

Might there be an economically, geographically and politically feasible route for the proposed canal, railway and oil pipeline that would entail significantly reduced risk? The general consensus in Nicaragua is no. Inhabitants — of all species — with ancient ties to the land will be uprooted regardless.

INTERNATIONAL ACTION

The contract for an interoceanic canal in Nicaragua represents a classic example of the challenges faced by a developing country in balancing economic growth and environmental protection. More sustainable ways to raise revenue and employment from Lake Nicaragua could include expanded irrigation, tourism and aquaculture. The population of Nicaragua is expected to grow by 37% by 2050, so water shortages and pressure on natural resources are already set to increase, limiting sustainable growth and public welfare. In preparation for a future of climate change, food insecurity and biodiversity loss, Nicaragua must establish long-term measures for the protection of its environment, not sacrifice itself to speculators.

A loose coalition of more than 30 concerned groups filed legal complaints with the government of Nicaragua in the second half of last year. These included three communities — the Miskitu and Ulwa indigenous peoples and the Rama–Kriol territorial government in the South Atlantic Autonomous Region — arguing that the canal concession violates their land rights and legal autonomy (see go.nature.com/ttshoc). These legal petitions were overridden by the National Assembly in December.

Swift and decisive international action is called for. The Nicaraguan Academy of Sciences (of which one of us, J.A.H.-P., is president) is coordinating efforts with the InterAmerican Network of Academies of Sciences to carry out an independent impact assessment. We need more conservation groups and social organizations to lend their expertise and funds if we are to prevent the tragic devastation of indigenous communities along with terrestrial, marine and freshwater biodiversity and resources in Central America. ■

Jorge A. Huete-Pérez *is director of the Centre for Molecular Biology at the Universidad Centroamericana, Managua, Nicaragua, and the president of the Nicaraguan Academy of Sciences.* **Axel Meyer** *is professor of zoology and evolutionary biology at the University of Konstanz, Germany.*
e-mails: jorgehuete@uca-cbm.org; axel.meyer@uni-konstanz.de

1. Comisión de Trabajo del Gran Canal *Interoceanic Canal Through Nicaragua: Project Profile* (2006); available at http://go.nature.com/psvowd (Spanish).
2. Klytchnikova, I. I., Cestti, R. E., Escurra, J., Jorge J. & Pagiola, S. P. *Policy and Investment Priorities to Reduce Environmental Degradation of the Lake Nicaragua Watershed (Cocibolca)* (World Bank, 2013); available at http://go.nature.com/91ruql.
3. Elmer, K. R., Kusche, H., Lehtonen, T. K. & Meyer, A. *Phil. Trans. R. Soc. B* **365,** 1763–1782 (2010).
4. Wilber, C. G. *Turbidity in the Aquatic Environment: An Environmental Factor in Fresh and Oceanic Waters* (Charles Thomas, 1983).
5. McKaye, K. R. *et al. BioScience* **45,** 406–411 (1995).
6. Barluenga, M., Stölting, K. N., Salzburger, W., Muschick, M. & Meyer, A. *Nature* **439,** 719–723 (2006).
7. Verones, F., Pfister, S. & Hellweg, S. *Environ. Sci. Technol.* **47,** 9799–9807 (2013).
8. Corbett, J. J. *et al. Environ. Sci. Technol.* **41,** 8512–8518 (2007).
9. Jordan, C. A., Stevens, K. J., Urquhart, G. R., Kramer, D. B. & Roe, K. *Tapir Conserv.* **19,** 11–15 (2010).

SCIENCE PHOTO LIBRARY

Clostridium difficile **(yellow cells) causes an intestinal infection that can be treated with processed stool.**

How to regulate faecal transplants

For medical use, human stool should be considered a tissue, not a drug, argue **Mark B. Smith**, **Colleen Kelly** and **Eric J. Alm**.

It is just over a year since the publication of the first randomized controlled trial[1] investigating the medical use of human faeces. The 43 trial participants had recurrent *Clostridium difficile* infections, which cause dangerous, painful and persistent diarrhoea. Those in the control groups received antibiotics alone. Those in the test group received antibiotics along with a fluid derived from filtered faeces, which was delivered into the upper small intestine through nasal tubes.

This small trial was stopped ahead of schedule because the faecal slurry was more than twice as effective in resolving symptoms as antibiotics alone[1]. Non-randomized studies, with outcomes collected from hundreds of people suffering from recurrent *C. difficile* infections and treated with similar procedures, have had typical success rates of around 90% (ref. 2).

First described[3] in the scientific literature in 1958, faecal microbiota transplantation (FMT), delivers processed stool from a healthy individual to the gut of a sick person through enema, colonoscopy or other means. The goal is to displace pathogenic microbes from the intestine by re-establishing a healthy microbial community. Interest has surged in the past five years (see 'Stool treatment'). At the same time, new regulatory barriers have made FMT more difficult to study or practice.

In May 2013, the US Food and Drug Administration (FDA) issued a public announcement that it had been regulating human faeces as a drug. This classification requires physicians to submit a time-consuming Investigational New Drug (IND) application before performing FMT. The FDA reasoned that this requirement would make FMT safer by providing oversight, standardizing therapy and, eventually, encouraging development of commercial drug products.

At a public meeting hosted that month by the FDA and the US National Institutes of Health (NIH), patients, physicians and representatives of the Centers for Disease Control and Prevention and several professional medical societies voiced concern about restricting access to care for these increasingly prevalent infections. Six weeks later, the FDA revised its position. The agency decided, for the time being, not to enforce the IND requirement for recurrent *C. difficile* infections.

This compassionate exception is now enabling many people to receive much-needed care. But the long-term status of FMT for *C. difficile* infection is unresolved, and regulatory policy is complicating research into the exploration of FMT for other conditions, such as inflammatory bowel diseases or obesity.

Whether and when the therapeutic potential of FMT is realized will depend on how the FDA and other agencies regulate the use of stool. Although treating it as a drug creates strict requirements to protect patients, it limits access to care. Reclassifying stool as a tissue product or giving it its own classification, as the FDA does for blood, would keep patients safe, ensure broad access and facilitate research.

RISKS AND BENEFITS

The human gut microbiome has been described as a 'virtual organ'[4]. Conditions ranging from inflammatory bowel diseases and obesity to asthma and cancer have been linked to its composition, with associations described between gut bacteria, biologically active metabolites and the immune system[5]. In addition to evidence from human studies of FMT, experimental evidence from studies in mice shows that changing this microbial ecosystem can affect hosts' physiology[6].

More than a half a dozen clinical trials have been registered to study FMT in inflammatory bowel diseases. But hopes that manipulating the gut microbiome to treat diseases other than *C. difficile* are still speculative[7,8].

Few human studies have followed patients prospectively to assess for adverse outcomes of FMT. Transient abdominal discomfort and bloating have been observed after FMT, but little long-term safety data exist.

Furthermore, there are real (albeit unrealized) risks that transplanting faecal microbiota can spread infectious diseases such as HIV or hepatitis. In the 1970s and 1980s, before today's strict blood-donation regulations were introduced, thousands of people with haemophilia in the United States were infected with HIV from contaminated blood products. There are also theoretical hazards that FMT could change the microbiome to make people more susceptible to chronic conditions such as obesity or autoimmune disorders. (The same can be said for the use of antibiotics, which might also cause unknown, lasting perturbations to the microbiome.)

Risks of FMT can be mitigated by mandating rigorous screening. But overly

restrictive rules might encourage people to seek treatment outside the medical establishment. Instructions for do-it-yourself faecal transplants are available online; individuals have posted videos on YouTube with tens of thousands of views and written books advocating at-home procedures using stool from acquaintances or family members. Some have even approached us for advice about using their pets as donors. An open letter on one FMT advocacy site urges doctors to recognize that at-home treatments are occurring, partly because physicians are not offering the procedure (see go.nature.com/zrzbuk).

The current situation is one of both under- and over-regulation. FMT for recurrent *C. difficile* infections can be performed without any mandatory screening, whereas FMT for other indications cannot be performed without an IND, a hurdle that will dissuade some physician–investigators.

The FDA defines drugs, in part, as "articles intended for use in the diagnosis, cure, mitigation, treatment, or prevention of disease". However, stool is unlike conventional drugs, which are produced under controlled conditions with consistent, known ingredients. Stool is a variable, complex mixture of microbes, metabolites and human cells. It cannot be characterized to the rigorous standards applied to conventional drugs. The material is also widely available — it comes from healthy volunteers, rather than chemical factories or controlled cell cultures.

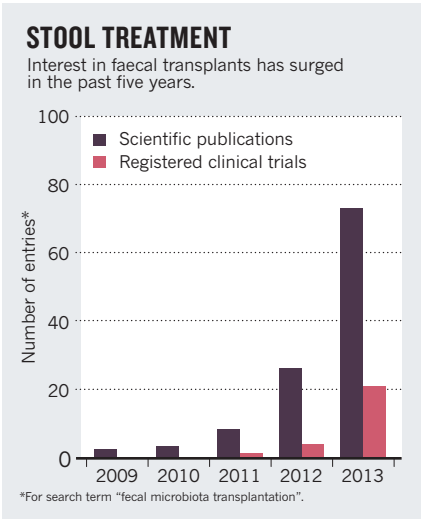
The FDA regulates blood, cartilage, bone, skin and egg cells as human tissues or under similar customized statutes. Transplanting these products into people requires meticulous record-keeping and screening for communicable diseases. These are exactly the safety precautions that should be taken with FMT. Under current US law, products that are excreted from the body or that depend on living cells from non-relatives are excluded from this category, disqualifying faeces. Exceptions have been made for semen, which is a tissue product, and blood, which has its own bespoke rules. In our view, statutes should be changed so that faeces too can be regulated as a tissue, not a drug. Risks can be reduced by rigorous screening, and the potential for clinical benefit is substantial.

STOOL BANKS

Appropriate regulation would pave the way for carefully screened and processed material to be made available through stool banks that operate similarly to blood banks. Stringent safety requirements set by the FDA would protect against infectious diseases, and a mandatory registry would track adverse events. Centralizing the screening and processing steps would make the treatment cheaper, safer, less variable and more convenient. It would also reduce the demand for risky at-home procedures.

As a model for this approach, we (M.B.S. and E.J.A.) helped to launch a stool bank called OpenBiome in 2012, which supplies material for *C. difficile* treatments, under the exemption currently offered by the FDA. This non-profit organization is funded primarily by charitable donations, but in the future, this and other stool banks could be sustained through user fees from hospitals. In its first three months of operation, OpenBiome delivered more than 100 treatments for *C. difficile* to 12 US hospitals. At least two teaching hospitals, including Massachusetts General Hospital in Boston and Emory University Hospital in Atlanta, Georgia, have also developed stool banks for their own patients.

OpenBiome screens donors for infectious agents through 17 blood and stool assays; donors are also assessed for chronic conditions, such as metabolic syndrome, autoimmune disorders and digestive problems. Many samples, often dozens, are collected from each donor, reducing screening costs per treatment to US$250, a fraction of the cost of one-off treatments. Samples are homogenized, filtered and frozen for long-term storage, providing physicians with a standardized, convenient source of material. This model could easily be scaled to meet the clinical need for FMT, both for *C. difficile* and for clinical trials of other diseases.



SOURCE: PUBMED; CLINICALTRIALS.GOV

Although some companies such Rebiotix in Roseville, Minnesota, and Monarch Laboratories in Irvine, California, are hoping to commercialize faeces-derived products as a drug, this classification threatens to restrict FMT mainly to companies with the resources to fund large clinical trials. Stool banks, which would distribute faeces as a tissue, could promote access and allow more investigation of potential clinical uses of FMT.

There are diseases in which the gut microbiome has been implicated but that do not have sufficient evidence to warrant FMT. To discourage inappropriate use, tissue banks should require documented approval from a clinic's institutional review board before releasing material for conditions other than *C. difficile* infection.

SYNTHETIC COMMUNITIES

In the past decade, our understanding of the microbiome has moved from identifying species to associating them with diseases. The next step is to engineer this system to improve human health. FMT, as it is currently practiced, does not use specific, pure cultures of bacterial isolates; rather, it delivers uncharacterized, minimally processed faecal material into a patient. As this field progresses, we expect that the microbial 'active ingredients' will be elucidated, enabling well-characterized cultures to be used as a second generation of microbiome therapeutics[9,10].

Knowledge gleaned from transplantation of natural systems could inform the design of synthetic communities tailored to treat specific diseases. Unlike stool, these derivatives will be easy to regulate as drugs because they will be highly characterized, consistently manufactured cocktails of bacteria and will not be freely available from a friend. Companies such as Seres Health in Cambridge, Massachusetts, and Vedanta Biosciences in Boston have already started testing this synthetic approach.

Nonetheless, treatments using synthetic communities remain several years off, and growing evidence suggests that FMT can effectively treat many patients now. Regulating faeces as a human tissue could help patients immediately and accelerate research into refined alternatives. ■

Mark B. Smith *is a PhD candidate in microbiology at the Massachusetts Institute of Technology in Cambridge, Massachusetts, and co-founder of OpenBiome.* **Colleen Kelly** *is a gastroenterologist and clinical assistant professor of medicine at the Alpert Medical School of Brown University in Providence, Rhode Island.* **Eric J. Alm** *is associate professor of biological engineering at the Massachusetts Institute of Technology.*
e-mail: ejalm@mit.edu

1. van Nood, E. *et al. N. Engl. J. Med.* **368,** 407–415 (2013).
2. Kassam, Z., Lee, C. H., Yuan, Y. & Hunt, R. H. *Am. J. Gastroenterol.* **108,** 500–508 (2013).
3. Eiseman, B., Silen, W., Bascom, G. S. & Kauvar, A. J. *Surgery* **44,** 854–859 (1958).
4. Evans, J. M., Morris, L. S. & Marchesi, J. R. *J. Endocrinol.* **218,** R37–R47 (2013).
5. Cho, I. & Blaser, M. J. *Nature Rev. Genet.* **13,** 260–270 (2012).
6. Riduara, V. K. *et al. Science* http://dx.doi.org/10.1126/science.1241214 (2013).
7. Smits, L. P., Bouter, K. E. C., de Vos, W. M., Borody, T. J. & Nieuwdorp, M. *Gastroenterology* **145,** 946–953 (2013).
8. van Nood, E., Speelman, P., Nieuwdorp, M. & Keller, J. *Curr. Opin. Gastroenterol.* **30,** 34–39 (2014).
9. Lawley, T. D. *et al. PLoS Pathog.* **8,** e1002995 (2012).
10. Petrof, E. O. *et al. Microbiome* **1,** 3 (2013).

ISSEI KATO/REUTERS/CORBIS

Journalists visit the Fukushima Daiichi nuclear power plant in Japan ahead of the first anniversary of its meltdown.

NUCLEAR ENERGY

Meltdowns, redux

Two accounts take contrasting lessons from nuclear accidents, finds **Mark Peplow**.

How safe is safe enough? Dig into the nuclear-power debate, and you will soon reach that question. Two books offer answers — but arrive at utterly different conclusions.

In *Atomic Accidents*, James Mahaffey tries to persuade us that the mighty atom is our friend by showing how much nuclear engineers (he is one) have learned from the industry's mistakes. Whereas he puts accidents under the microscope to pinpoint where things turned nasty, in *Fukushima*, David Lochbaum, Edwin Lyman and Susan Q. Stranahan blame the entire nuclear establishment.

Mahaffey guides us through more than a century of atomic research, including misadventures with radioactive elixirs ("The radium water worked fine until his jaw came off", reads a 1932 headline) and long-forgotten accidents at enrichment plants. Along with show-stoppers such as Three Mile Island and Chernobyl, he covers milestones including the first weapons-test accident (in 1954, when the detonation of the compact US hydrogen bomb 'Shrimp' unintentionally contaminated 18,000 square kilometres of the Pacific Ocean) and the first meltdown (1952, at Canada's Chalk River reactor). The accidents are mostly reconstructed from official reports, and Mahaffey includes a lot of technical detail that serves as a useful introduction to nuclear engineering.

Atomic Accidents: A History of Nuclear Meltdowns and Disasters from the Ozark Mountains to Fukushima
JAMES MAHAFFEY
Pegasus Books: 2014.

Fukushima: The Story of a Nuclear Disaster
DAVID LOCHBAUM, EDWIN LYMAN AND SUSAN Q. STRANAHAN
The New Press: 2014.

Entertaining anecdotes about foolhardy pioneers abound. In the 1940s, after diving into a spent-fuel pool to adjust an experiment, bomb-core-assembly expert Louis Slotin was moved from Oak Ridge National Laboratory in Tennessee to Los Alamos National Laboratory in New Mexico, "where daring was better appreciated". He was killed 18 months later, in a stupid accident with a screwdriver and some plutonium. The compelling tales unravel like slow-motion horror stories, spiralling towards disasters we know are coming.

A theme emerges. Accidents happen when operators do not follow the correct procedures, or because ambitious plant designers overlook glaring weaknesses — not because nuclear power is inherently unsafe. The disaster at Japan's Fukushima Daiichi power plant — triggered by the 2011 earthquake and tsunami — is afforded fewer than 30 pages. As with his other accounts, Mahaffey quickly identifies the accident's turning points. In the case of the plant's Unit 1 reactor, which suffered a complete meltdown, he singles out an operator who closed two crucial coolant valves, effectively overriding an automated safety system.

He takes the same approach to the whole industry, picking a little-known US accident in 1961 as the moment that led plant designers to take a wrong turn. When a control rod was inadvertently pulled from SL-1, a low-power military reactor in remote Idaho, it caused a steam explosion that took the lives of three people — the last to die in a power-reactor accident in the United States. The incident soured the industry on small, simple reactors, and pushed it towards bigger, more expensive ones that became ever more complicated as safety features were retrofitted.

Mahaffey argues for a return to smaller reactors, reasoning that accidents are inevitable, so they had best be small. He extols the

NATURE.COM
For more on nuclear history, see:
go.nature.com/5zj6qk

virtues of safer designs such as the thorium molten-salt reactor. If these changes are made and lessons are learned, he concludes, accidents like Fukushima should be behind us.

On the contrary, say Lochbaum, Lyman and Stranahan. "Nuclear power is an energy choice that gambles with disaster," they write. "The problems that led to the disaster at Fukushima Daiichi exist wherever reactors operate." They unpick those problems in forensic detail, using multiple sources in a thriller-paced retelling. *Fukushima* takes a much broader view of the accident than *Atomic Accidents*, delving into political wrangling and the roles of international agencies. It shows how Japan's complex nuclear bureaucracy — involving power companies, an independent regulator and government departments — stymied the response. A vivid picture emerges of utter confusion in the hours and days after the tsunami.

The writers have impressive pedigrees. Stranahan was on the *Philadelphia Inquirer* team that won a Pulitzer Prize in 1980 for its coverage of the Three Mile Island accident. Industry insider Lochbaum and global-security specialist Lyman have both been heavily involved in the Union of Concerned Scientists' lobbying on nuclear power.

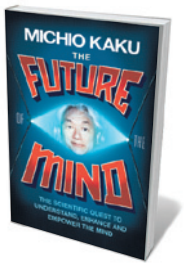
That may explain why the second half of the book becomes an attack on the US Nuclear Regulatory Commission (NRC), which the authors argue is complicit in the industry's disregard for safety. According to *Fukushima*, the NRC refused to learn from Three Mile Island, and failed to mandate that the industry prepare for similar events. The commission, the book claims, had run simulations showing that Mark 1 boiling-water reactors, designed by General Electric and installed at Fukushima, were vulnerable to meltdown in a power blackout. If the NRC had been bolder about improving safety at home, in the authors' opinion, other countries would have followed — and Japan might not be facing a US$100-billion nuclear clean-up.

Lochbaum, Lyman and Stranahan disagree strongly with Mahaffey's stance on the benefits of smaller reactors, which would almost certainly be built in clusters: at Fukushima, simultaneous problems with multiple reactors complicated emergency-response efforts. "Nuclear power's safety problems cannot be solved through good design alone," they write. Instead, they say, the NRC must accept the possibility that dam breaches, fires or terrorist attacks could trigger a nuclear accident worse than Fukushima on US soil.

Both polemics offer thought-provoking analyses. However much they differ, they are both right: if nuclear power is to have a future, it needs better science and better regulation. ■

Mark Peplow *is a science journalist based in Cambridge, UK.*
e-mail: peplowscience@gmail.com

Books in brief



The Future of the Mind: The Scientific Quest to Understand, Enhance, and Empower the Mind
Michio Kaku DOUBLEDAY *(2014)*
Taking a break from contemplating the cosmos, Michio Kaku plunges into the universe inside the skull, training his theoretical physicist's eye on the field. His intriguing 'space-time' theory of consciousness frames the extraordinary findings emerging from ever-more-finely targeted brain scanning and other technologies. A fascinating sprint through everything from telepathy research to the 147,456 processors of the Blue Gene computer, which has been used to simulate 4.5% of the brain's synapses and neurons.

Girls Coming to Tech! A History of American Engineering Education for Women
Amy Sue Bix MIT PRESS *(2014)*
The Second World War flung open windows of opportunity on aircraft engineering for thousands of women in the United States. In the 1950s, many eager to pursue an engineering degree hit a wall; less than 1% of the era's engineering students were female. Focusing on three iconic technology institutes (California, Georgia and Massachusetts), science historian Amy Sue Bix relates how these "oddities at best and outcasts at worst" made headway in closing the gender gap: women now earn one-fifth of degrees in the field.

The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies
Erik Brynjolfsson and Andrew McAfee W. W. NORTON *(2014)*
In this comparative study of economic and digital progress, Erik Brynjolfsson and Andrew McAfee argue that we stand at an "inflection point" — poised to reap big rewards if we harness the forward leap of innovation. With measured optimism, they survey a digital landscape of exponential progress in computing power and application; technological benefits and their uneven spread; and policy. Crammed with analyses of everything from human–machine competition to the state of US education.

Joy, Guilt, Anger, Love: What Neuroscience Can — and Can't — Tell Us About How We Feel
Giovanni Frazzetto PENGUIN BOOKS *(2014)*
Neuroscientist Giovanni Frazzetto enters the restless realm of human emotion through the portals of physiology, genetics, history, art and philosophy. Anger, guilt, anxiety, grief, empathy, joy and love are anatomized in turn, enlivened with research on everything from the role of monoamine oxidase A in anger to the engagement of opioid receptors as we thrill to music. And who knew that surrealist Salvador Dali created an art installation in the shape of a giant caterpillar to explore the process of sedation?

GDP: A Brief but Affectionate History
Diane Coyle PRINCETON UNIVERSITY PRESS *(2014)*
A raft of economists, including Robert Costanza (see *Nature* **505,** 283–285; 2014), argue that gross domestic product (GDP) is a flawed measure of national prosperity, hiding social inequality and pushing growth at the planet's expense. Economist Diane Coyle is less severe in this brief, lucid history. She traces GDP from its roots in the eighteenth century to its twentieth-century heyday, offering a smart analysis of its status and uses now, as a one-note statistic in an increasingly complex world. **Barbara Kiser**

A toy replica of NASA's failed Mars Climate Orbiter is now a collector's item.

SCIENCEGALLERY.COM

RESEARCH AND DEVELOPMENT

No matter, try again

Anthony King enjoys a collection of instructive failures.

An exhibition exhorts us to embrace failure. Each of the 20 objects displayed in *Fail Better* at the Science Gallery, Trinity College Dublin, has been chosen by a luminary to represent an inspiring or arresting story of the role of failure in stimulating creativity.

Inspired partly by the mantra of start-up entrepreneurs — "Fail early, fail fast, fail often" — the show also recalls the words of Trinity College Dublin graduate Samuel Beckett: "Ever tried. Ever failed. No matter. Try again. Fail better." He saw failure as the ultimate goal of art.

Science, too, is built from mistakes that are, as Jules Verne wrote, "useful to make, because they lead little by little to the truth". For example, astrophysicist Jocelyn Bell Burnell's pick for the exhibition is the Mars Climate Orbiter, the US$125-million spacecraft that went missing in September 1999. A week later, NASA announced: "people sometimes make errors". Notoriously, one team had used metric units, the other imperial — a simple mistake that probably led to the orbiter disintegrating in the Martian atmosphere.

Fail Better
Science Gallery, Trinity College Dublin. Until 27 April 2014.

Says Bell Burnell's programme note: "Even at the highest level errors happen and can go unnoticed, proving that attention to detail is always paramount." On show is a toy replica of the orbiter, which was hastily withdrawn by NASA and is now a collector's item. The loss prompted NASA to look anew at its 'faster, better, cheaper' approach.

By contrast, to signify how it is important to fail on a small scale rather than spectacularly, there is a humble fuse — a reflection by economist Tim Harford on how our economic system lacked such a 'fuse' to prevent the 2008 financial meltdown.

The most bizarre exhibit is a mock up of a Blonsky device. Granted a patent in 1965 but never built, the machine was designed to spin a woman during childbirth at up to 7*g* to assist delivery. It features a baby-blue 'safety net' to catch the new arrival. The choice of Ig Nobel Prize founder Marc Abrahams, the device won inventors George and Charlotte Blonsky one of the gongs in 1999.

That mistakes can be costly and destructive is rammed home by *The Ice Pick Lobotomy*, an installation by Trinity neuroscientist Shane O'Mara. One of several medical instruments on show, the macabre lobotomy kit from the Wellcome Collection of medical artefacts reminds us of a terrible, telling failure of medical ethics. But it is a shame that the exhibition does not delve more into the issue of hidden failure in research. Too often, the messiness of experimentation is obscured, and the lessons of failure squandered, thanks to the bias towards publishing successful results.

"Too often, the messiness of experimentation is obscured, and the lessons of failure squandered."

What we do see is the role of serendipity in discovery, and the importance of persistence. A structure sculpted from purple acrylic thread embodies the accidental invention of the colour mauve, when a botched 1856 attempt by chemist William Perkin to make quinine from coal tar ended up founding the synthetic-dye industry. Meanwhile, a cutaway Dyson vacuum cleaner is described as the result of more than 2,000 prototypes by its selector, James Dyson.

Finally, there are three manuscript drafts of Beckett's *Worstward Ho*, the 1983 novella in which he lauded failure. Each uses different iterations of the line quoted here, as Beckett made revisions. He reputedly railed against winning the Nobel Prize in Literature, fearing that the publicity would distract him from his writing. He knew, it seems, that the path to success through a thicket of frustration is often the more instructive journey. ■

Anthony King *is a writer based in Dublin.*
e-mail: anthonyjking@gmail.com

CORRECTION
The Books & Arts Q&A 'Tinnitus tunesmith' (*Nature* **505,** 159; 2014) omitted the name of neuroscientist Tricia MacKenzie, who likened Daniel Fishkin's instrument to a giant model of the inner ear.

Correspondence

Move on to a carbon currency standard

Alongside Robert Costanza and colleagues' plea to abandon gross domestic product as a measure of national success (see *Nature* **505,** 283–285; 2014), we believe that there is an urgent need to change the way currencies are valued — by using a new 'carbon standard' that links economy to ecology.

This would work in a similar way to the old gold-exchange standard, except that a country's currency value would instead be determined by its saved and standing stocks of fossil and non-fossil carbon. Governments would need to decide whether to risk devaluing their currency by depleting carbon stocks — while still honouring a commitment to keep fossil-carbon stocks at 80% as a safeguard against extreme climate change.

After the Second World War, huge investments radically altered the economies of the United States, the Soviet Union and the United Kingdom. In the face of climate change, it is now the global energy system that needs reinvention.

John R. Porter *University of Copenhagen, Denmark.*
jrp@plen.ku.dk
Steve Wratten *Lincoln University, New Zealand.*

Apply market forces to peer review

Dan Graur suggests that journals should devise a system to tie in the number of papers they publish from senior researchers with the reviewing record of those authors (*Nature* **505,** 483; 2014). This is comparable to linking the amount of milk you drink to the number of cows you milked.

In a market economy, the monetary value of services is determined by the laws of supply and demand. When it comes to the highly skilled service of peer reviewing, the supply is sufficiently high to keep the monetary value at zero.

If, at a constant level of demand, the supply is reduced, then this price would go up. With an increased price, people could become professional reviewers to supplement their salary. Instead of making it harder for scientists to decline their reviewing requests, journals should be allowing market forces to exert their natural effect.

Sascha Ott, Daniel Hebenstreit *University of Warwick, UK.*
s.ott@warwick.ac.uk

Drug-resistant TB can be contained

The spread of multi-drug-resistant tuberculosis (MDR-TB) in Russia is by no means unstoppable (*Nature* http://doi.org/rf3; 2014). MDR-TB epidemics have been prevented and/or reversed in Latvia, Estonia, Hong Kong and the United States, and even in the hard-hit Russian provinces of Orël and Tomsk, by using evidence-based policies and patient-centred care.

Containing the airborne spread of MDR-TB requires transmission to be interrupted, treatment of active disease with the right antibiotics for the correct amount of time, treatment of latent infection, and locating cases and their contacts.

As a result of political determination and a series of intervention programmes, the Tomsk Oblast TB services, in conjunction with Partners In Health, based in Boston, Massachusetts, reduced MDR-TB prevalence from an estimated 823 cases in 2002 to 328 in 2012. These reductions were the result of structural changes (better facilities, treatment sites and community-based care), clinical improvements (individualized drug regimens and aggressive management of adverse events) and social interventions (treatment incentives and assistance with food and transportation).

The complexity and expense of MDR-TB treatment take a huge toll on patients and carers. We urgently need better diagnostics, shorter and less-toxic treatments, and more research and resources. Even with existing resources, much more is possible.

Erica Lessem *Treatment Action Group, New York, USA.*
erica.lessem@treatmentactiongroup.org
Salmaan Keshavjee *Harvard Medical School, Boston, Massachusetts, USA.*

Sci-fi should not discredit vaccines

Ordinarily, I would not consider a deviation from absolute scientific truth to be a problem in science fiction. But I am disturbed by the dangerous and incorrect implications in A. G. Carpenter's Futures story that vaccines weaken the human immune system, and that their effects are on a par with poisoning by lead or tobacco (*Nature* **506,** 126; 2014).

Vaccines are safe, effective and rigorously tested (see, for example, *Nature Immunol.* **9,** 1317; 2008). Ungrounded fears already prompt parents unnecessarily to opt out of vaccination programmes for their children, leading to entirely avoidable illnesses and diminishing herd immunity.

In my view, it is irresponsible to promote the idea that vaccines are bad for health — even fictitiously.

Rachel Reddick *Stanford University, California, USA.*
rmredd@stanford.edu

NATURE'S READERS COMMENT ONLINE

A selection of views on the topic of senior scientists turning down review requests (D. Graur Nature ***505,*** *483; 2014).*

Paul Flicek says:
Graur makes the anecdotal observation that those scientists who publish most often are more likely to turn down requests to review papers. But manuscripts are only a part of a senior scientist's peer-reviewing activities: these include peer review for individuals and institutions, and of many more grant applications than they themselves submit.

S. Sudha Tushara and S. Sudarsan say:
A reviewing 'impact factor' awarded to reviewers per paper would entice reviewers, and help editorial boards and funding agencies.

Bob O'Hara says:
Perhaps scientists who publish more are asked to review more, and thus even if they review as many manuscripts as everyone else, they will still appear to be less cooperative because they decline more manuscripts.

S. A. Luis, M. Madadin and R. G. Menezes say:
Nominal financial incentives would help too, as would mandating involvement with peer review as a prerequisite for academic promotion.

Samad E. J. Golzari says:
The quality of a mandatory peer review might not be the same as a review undertaken voluntarily.

ASTROPHYSICS

Lopsided stellar death

Observations of the supernova remnant Cassiopeia A offer an unprecedented look back at the centre of this explosion, and support the hypothesis that spatial asymmetry is key to explaining the event. SEE LETTER P.339

J. MARTIN LAMING

Elements of the periodic table from carbon upwards are synthesized in massive stars during their evolution, and also in the supernovae by which such stars end their lives. Following these explosive events, the stars' nucleosynthetic products are redistributed back to the interstellar medium from which the stars formed. Partial validation of these ideas has come from observations of the abundances of elements in the Solar System, which necessarily give results averaged over many supernovae. Detailed examination of such processes in individual supernovae, especially those occurring at their centres, where iron-group elements are produced, has remained intractable. On page 339 of this issue, Grefenstette *et al.*[1] present the first measurement of the spatial distribution of the nuclear-decay products of titanium-44, a radioactive nucleus produced at the heart of such explosions. The result potentially heralds a new phase in our understanding of these events.

Stars that are born with more than about eight times the Sun's mass end their lives as core-collapse supernovae. When the energy to power the star from nuclear-fusion reactions is exhausted, which occurs when nuclear reactions in the core have proceeded as far as fusing nickel-56 (^{56}Ni), thermal pressure can no longer support the star's core against its gravity. Once the ^{56}Ni core has grown large enough, it collapses until it has the density of an atomic nucleus and stiffens. Further infalling material from the overlying stellar envelope can bounce off the collapsed core, and, if sufficiently energetic, the shock thus produced can cause the star to explode. This class of explosion is quite different from the thermonuclear incineration of a white-dwarf star, which results in the type Ia supernovae that are used to trace the Universe's expansion.

In fact, discovering a robust mechanism to explain core-collapse explosions has proved elusive. Most of the energy liberated in the collapse is radiated as neutrinos, which couple only weakly to matter and do not reliably re-energize the shock. The most promising ideas for producing a successful explosion seem to require strong departures from spherically symmetrical events[2–4].

Set against this background, Grefenstette and colleagues' study of the young, nearby supernova remnant Cassiopeia A (Cas A) represents a landmark. Using NASA's Nuclear Spectroscopic Telescope Array (NuSTAR), a space-based X-ray observatory, the authors recorded a map of the emission associated with the decay of titanium-44 (^{44}Ti), which is produced in the inner regions of the core-collapse explosion. They find that this map indeed shows strong departures from spherical symmetry, supporting the idea that asymmetry is key to explaining the explosion, and in particular that low-mode (dipole or quadrupole) oscillations of the core are crucial. But this map contains another surprise. It had been thought[5] that ^{44}Ti is always produced co-spatially with ^{56}Ni, which decays radioactively to cobalt-56 and then to iron-56 (^{56}Fe). However, the ^{44}Ti as imaged by NuSTAR is spatially distributed quite differently from ^{56}Fe as observed by X-ray and infrared instruments on the Chandra[6] and Spitzer[7] satellites, respectively.

In April–May 2004, the Chandra X-ray satellite observed Cas A for 1 million seconds, as one of the first of its 'Very Large Projects'. Detailed analysis[6] of these data found X-rays from Fe ions, but the ions were located towards the outer regions of the remnant and not in the centre as ^{44}Ti is (Fig. 1). The analysis also found various indications of pure Fe, but these 'ashes', which presumably include trace amounts of ^{44}Ti, were again found towards the remnant's periphery. In fact, the central ejecta of Cas A have not yet been heated by the reverse shock

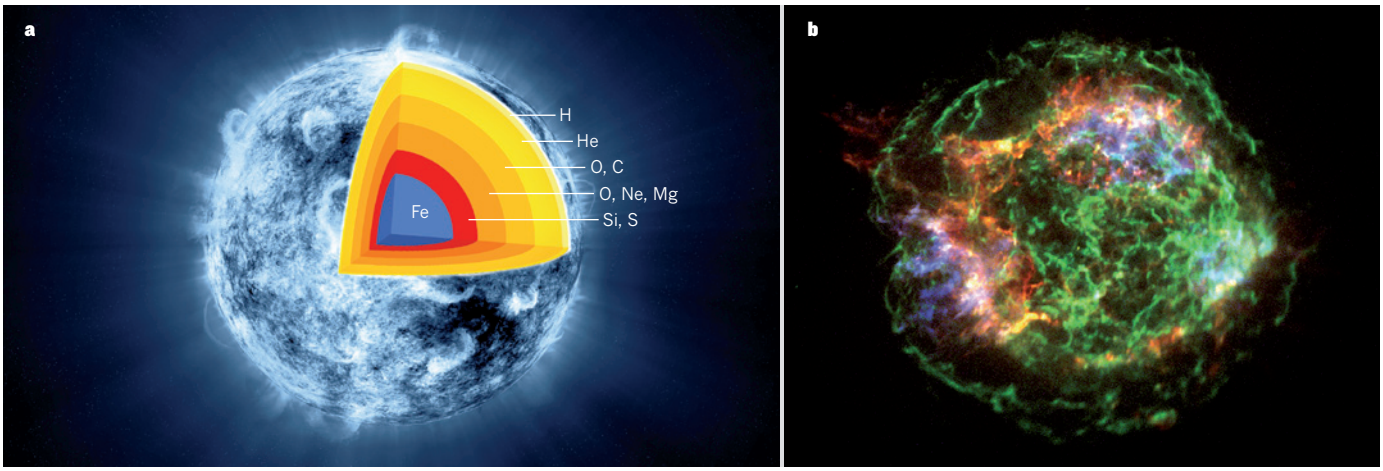


Figure 1 | A dying star turned inside out. **a**, Artist's reconstruction of the stratification of element composition in the Cassiopeia A progenitor star just before it exploded as a supernova. **b**, The supernova remnant as seen today in X-ray data, with the observed distributions of iron (blue) and silicon (red). The other elements are omitted for clarity, with synchrotron radiation from cosmic-ray electrons shown in green. The star seems to have turned 'inside out': the iron lies outside the silicon, especially on the left-hand side. According to current ideas, the iron should have been produced co-spatially with ^{44}Ti, but Grefenstette *et al.*[1] find ^{44}Ti towards the centre of the remnant (not shown). (**a**: Adapted from an illustration by NASA/CXC/M. Weiss; **b**: NASA/CXC/GSFC/U. Hwang *et al.*[10].)

that propagates inwards through the supernova material, and therefore emit primarily in the infrared region of the electromagnetic spectrum. Observations obtained with the Spitzer infrared satellite show no indication so far of Fe in the unshocked centre[7].

These and other analyses of X-ray and infrared data are far from straightforward. They involve calculations of the distribution of Fe between its various charge states, which in turn include assumptions about the remnant's plasma density and the time evolution of its shock waves. But even given these uncertainties, it seems unlikely that the conclusion that most of the ^{56}Fe is in the outer regions of Cas A, with the ^{44}Ti occurring in the centre, will be overturned.

Where does this leave us? All established models of core-collapse explosions predict ^{56}Fe existing co-spatially with ^{44}Ti, and it is difficult to see a way out of this problem unless a mechanism for decoupling the synthesis of ^{44}Ti and ^{56}Ni can be found. Despite being an energetic explosion and ejecting a significant mass of ^{44}Ti, Cas A is one of the historical supernovae that was apparently not observed at the time of its explosion[8], around AD 1670–80. This has led to speculation that the neutron star that formed at the centre of the remnant experienced a transition to an exotic type of star (a quark star) by undergoing a second explosion a few days after the supernova itself[9]. The second explosion (a quark nova) would have produced a wind of high-speed neutrons that would have broken up ^{56}Ni in the inner ejecta and enhanced the abundance of ^{44}Ti. But this means of decoupling the synthesis of ^{44}Ti and ^{56}Ni is controversial, and Cas A offers no more than a tantalizing hint about it.

It remains to be seen whether other explanations can be made to work. NuSTAR will continue to collect data from Cas A to improve the precision of the ^{44}Ti map, and further analysis, particularly of the Spitzer infrared data, is required to place a firm limit on the mass of ^{56}Fe that may reside in the inner ejecta. ■

J. Martin Laming *is in the Space Science Division, Naval Research Laboratory, Washington DC 20375, USA.*
e-mail: laming@nrl.navy.mil

1. Grefenstette, B. W. *et al. Nature* **506,** 339–342 (2014).
2. Foglizzo, T., Galletti, P., Scheck, L. & Janka, H.-Th. *Astrophys. J.* **654,** 1006–1021 (2007).
3. Scheck, L., Janka, H.-Th., Foglizzo, T. & Kifonidis, K. *Astron. Astrophys.* **477,** 931–952 (2008).
4. Brandt, T. D., Burrows, A., Ott, C. D. & Livne, E. *Astrophys. J.* **728,** 8 (2011).
5. Arnett, D. *Supernovae and Nucleosynthesis* (Princeton Univ. Press, 1996).
6. Hwang, U. & Laming, J. M. *Astrophys. J.* **746,** 130 (2012).
7. DeLaney, T. *et al. Astrophys. J.* **725,** 2038–2058 (2010).
8. Stephenson, F. R. & Green, D. A. *Historical Supernovae and their Remnants* (Oxford Univ. Press, 2002).
9. Ouyed, R., Leahy, D., Ouyed, A. & Jaikumar, P. *Phys. Rev. Lett.* **107,** 151103 (2011).
10. Hwang, U. *et al. Astrophys. J.* **615,** L117–L120 (2004).

REGENERATIVE BIOLOGY

Take the brakes off for liver repair

A protein produced by endothelial cells that line blood vessels has been found to regulate the timing of cell proliferation following liver injury, further demonstrating the role of vascular signals in tissue regeneration.

ANDREW G. COX & WOLFRAM GOESSLING

The liver's regenerative potential is legendary, and depends on a carefully orchestrated symphony of factors that enable a precise and timely recovery of the liver's metabolic and synthetic functions. The most frequently used model of liver injury is surgical removal of part of the liver, because this allows the regenerative process to be studied without associated inflammation and necrosis. Typically, the vertebrate liver recovers from partial hepatectomy in 5–7 days, and this seems to be the critical time frame within which hepatic function must be regained to enable survival. But how the liver senses injury and initiates and terminates regenerative growth is not fully understood. A role for hepatic blood vessels in regeneration was suggested four decades ago[1]. Now, writing in *Science*, Hu *et al.*[2] clarify that role by defining how an angiocrine factor — a signalling molecule produced by the endothelial cells that line the inner surface of blood vessels — is involved in the temporal control of the regenerative process.

The timeline of events after partial hepatectomy is well known and can be divided into two phases: during the inductive phase, which lasts between 1 and 3 days, proliferation of hepatocytes, the main liver cells, results in increased hepatocyte numbers; and during the subsequent angiogenic phase (referring to the formation of new blood vessels that occurs), all other liver-cell types proliferate to reconstitute the total liver mass.

By studying the transcriptome — the inventory of RNA molecules — of mouse liver sinusoidal endothelial cells (LSECs), Hu *et al.* discovered that the cells' expression of angiopoietin 2 (Ang2), a protein that promotes angiogenesis, is dynamically regulated during this process (Fig. 1). During the inductive phase, Ang2 expression decreases, leading

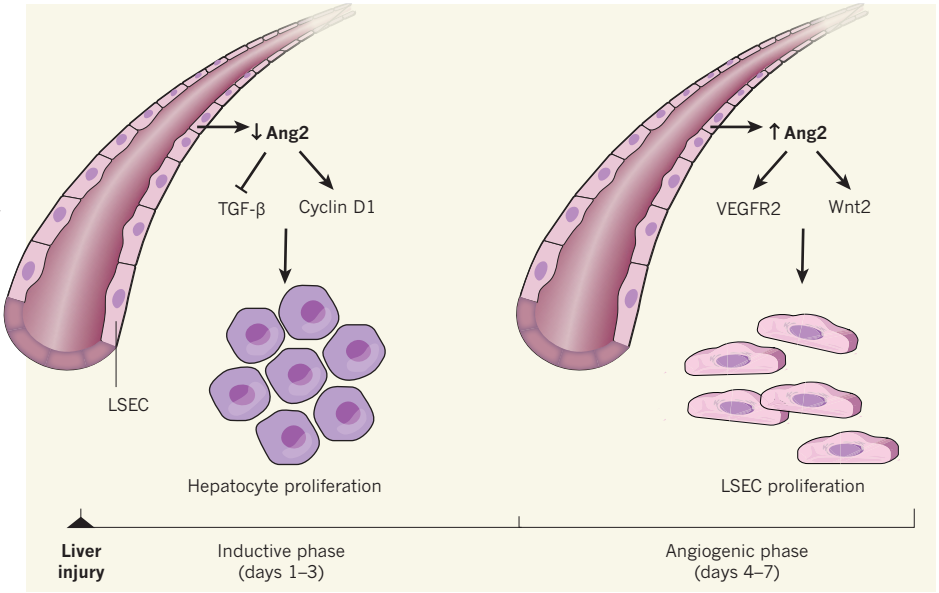


Figure 1 | Dynamic regulation of angiopoietin 2. Hu *et al.*[2] show that temporally regulated expression of the protein angiopoietin 2 (Ang2), which is expressed by liver sinusoidal endothelial cells (LSECs) lining the blood vessels, contributes to the two-phase process of liver repair following injury. During the inductive phase, low Ang2 expression causes reduced expression of transforming growth factor-β (TGF-β) and enhanced expression of cyclin D1; these effects combine to enhance the proliferation of hepatocytes, the primary liver cells. During the subsequent angiogenic phase, Ang2 expression rises, leading to increased expression of vascular endothelial growth factor receptor 2 (VEGFR2) and Wnt2, and thereby promoting the proliferation of LSECs.

to lower levels of a proliferation-inhibiting protein known as transforming growth factor-β and enhanced expression of the protein cyclin D1, culminating in accelerated hepatocyte proliferation. Then, during the angiogenic phase, Ang2 levels recover and stimulate the proliferation of LSECs in a process that depends on the vascular endothelial growth factor receptor 2 (VEGFR2) and on expression of the protein Wnt2.

The authors also show that Ang2 plays an important part in the response to chronic chemical injury following exposure of the liver to the toxin carbon tetrachloride. Thus, it seems that LSECs first act 'altruistically', to enhance hepatocyte recovery, and later boost their own growth. Interestingly, a report[3] measuring Ang2 levels in patients with acute liver failure revealed that higher levels are associated with a worse clinical outcome, further documenting the clinical and therapeutic relevance of Ang2 signalling and of the blood vessels that produce this protein.

An earlier study[4] had already demonstrated the importance of LSECs and VEGFR2 in liver repair, although in that work VEGFR2 was implicated in the inductive phase by mediating the expression of the proteins Id1, Wnt2 and hepatocyte growth factor (HGF), which contribute to the induction of hepatocyte proliferation. The same research group also showed[5] that angiocrine signalling controls the balance between optimal organ repair after acute injury and scarring after prolonged insult — expression of the receptor protein CXCR7 in LSECs is responsible for the pro-proliferative Id1–Wnt2/HGF response after acute injury. By contrast, CXCR7 expression is suppressed during chronic injury, leading to defective repair and scarring.

Future studies using cell-specific inactivation of Ang2 and VEGFR2 may help to clarify the complex interplay of these factors during the inductive and angiogenic phases of liver regeneration. However, Hu and colleagues' findings clearly show that Ang2 expression orchestrates the proliferation of hepatocytes and LSECs during this process. The angiogenic action of Ang2 is consistent with the protein's reported role in protecting blood vessels against stress[6]. Other angiocrine factors produced by the endothelium have been identified as contributors to liver regeneration, including prostaglandin E2 (ref. 7), epoxyeicosatrienoic acids[8] and nitric oxide[9]. The involvement of endothelial cells and their signals has also been demonstrated for proper regeneration of the lung[8,10], haematopoietic stem cells[11] and the kidney[8].

The liver responds swiftly to organ injury. But is it plausible that liver endothelial cells initiate and orchestrate this response? It seems so. After partial hepatectomy, the remaining liver tissue is intact and uninjured, and is not exposed to toxins or dying hepatocytes. By contrast, the vasculature is subject to immediate changes in blood flow and thus potentially to altered exposure to soluble signalling factors. Because the entire blood flow entering the liver circulates through a much reduced liver-cell mass, physical blood-flow parameters or soluble-factor concentrations will also change; this could represent the initial signal of injury immediately sensed by LSECs[1]. Similarly, liver injury by toxins leads to cell swelling and cell death, which alter blood-flow dynamics. Endothelial cells are therefore extremely well equipped to both sense and immediately respond to changes in the integrity, size and metabolic capacity of the liver.

One could also imagine a direct role for endothelial cells in promoting the proliferation of other liver-cell types, such as biliary epithelial cells or stellate cells, and in the cessation of proliferation once hepatic repair is complete. Is a single molecule responsible for regulating the entire process? Probably not, because optimal liver regeneration after injury is essential for the survival of the organism, and a multitude of factors and signals are known to act in concert, and redundantly, to achieve rapid and efficient functioning of the organ. Hu and colleagues' work, however, demonstrates that signals emerging from the hepatic vasculature are dynamically modulated to govern the entire temporal sequence of hepatic repair. What can we learn from this? That to promote hepatic regeneration in our sickest patients, we need to take the brakes off as much as we need to press the accelerator. ■

Andrew G. Cox *and* **Wolfram Goessling** *are in the Genetics and Gastroenterology Divisions, Brigham and Women's Hospital, Harvard Medical School, Boston, 02115 Massachusetts, USA.* **W.G.** *is also at the Dana-Farber Cancer Institute, the Harvard Stem Cell Institute and the Broad Institute of MIT and Harvard.*
e-mail: wgoessling@partners.org

1. Fisher, B., Szuch, P., Levine, M. & Fisher, E. R. *Science* **171,** 575–577 (1971).
2. Hu, J. *et al. Science* **343,** 416–419 (2014).
3. Hadem, J. *et al. Crit. Care Med.* **40,** 1499–1505 (2012).
4. Ding, B.-S. *et al. Nature* **468,** 310–315 (2010).
5. Ding, B.-S. *et al. Nature* **505,** 97–102 (2014).
6. Daly, C. *et al. Proc. Natl Acad. Sci. USA* **103,** 15491–15496 (2006).
7. North, T. E. *et al. Proc. Natl Acad. Sci. USA* **107,** 17315–17320 (2010).
8. Panigrahy, D. *et al. Proc. Natl Acad. Sci. USA* **110,** 13528–13533 (2013).
9. Cox, A. G. *et al. Cell Rep.* **6,** 56–69 (2014).
10. Ding, B.-S. *et al. Cell* **147,** 539–553 (2011).
11. Kobayashi, H. *et al. Nature Cell Biol.* **12,** 1046–1056 (2010).

CANCER

Persistence of leukaemic ancestors

The early development of acute leukaemias is assumed for the most part to be clinically silent and transient. But it now seems that ancestral precancerous cells are identifiable and persistent. SEE ARTICLE P.328

NICOLA E. POTTER & MEL GREAVES

Aggressive leukaemias often present clinically out of the blue, without previous indications of cancer. But evolutionary models of cancer development posit a time-ordered, stepwise process involving the accumulation of mutations, proliferation of mutated cells into expanded clonal populations and selection of the fittest cells[1]. Such models imply that any seemingly sudden case of cancer will have arisen from 'silent' precursor cells that have no clinical impact. In this issue, Shlush *et al.*[2] (page 328) provide compelling evidence that the early-stage cells of acute myeloid leukaemia are not completely outcompeted and rendered extinct by their more aggressively cancerous and numerous progeny, but instead persist and show defined genetic and functional properties.

The cancerous myeloid cells (a subset of white blood cells) of patients with acute myeloid leukaemia (AML) often have a mutation in the gene encoding the DNA-methyltransferase enzyme DNMT3a. In two studies of patients with *DNMT3a*-mutated AML — one involving 4 patients and the other 17 — Shlush and colleagues made the unexpected observation that, in 15 patients, the *DNMT3a* gene carried the same mutation at a low rate in T cells (a type of immune cell that belongs to the lymphoid system) in the blood. Strikingly, however, the T cells lacked other alterations present in the leukaemic cells of the same patients, including mutations in the gene *NPM1*. The *DNMT3a* mutation was also detected at variable frequency in other immune cells (B and NK cells) at the time of AML diagnosis. Finding this mutation in non-myeloid cells hinted at its occurrence in a precursor cell that gives rise to blood cells of both the lymphoid and myeloid lineages.

The frequency with which *DNMT3a* and *NPM1* were mutated in leukaemic cells was

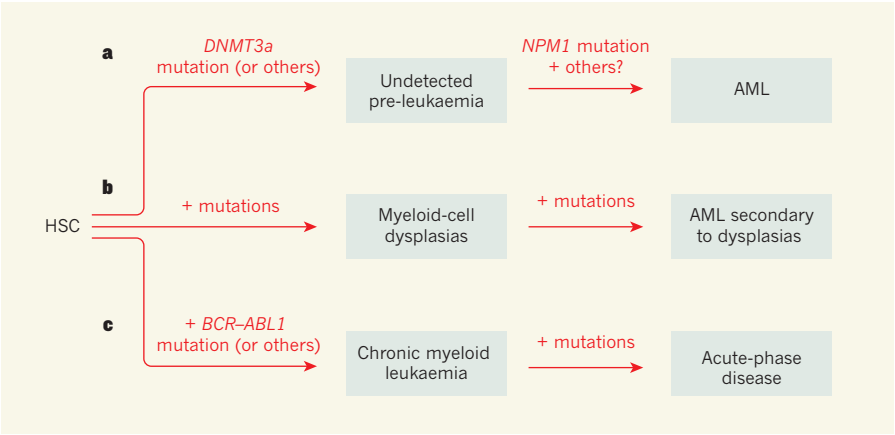


Figure 1 | Different routes to acute myeloid leukaemia. Cancer occurs when cells accumulate mutations over time. In acute myeloid leukaemia (AML), the first cell to be transformed to a cancer-like state is typically a haematopoietic stem cell (HSC). Differentiation of this cell can lead to AML through one of three intermediate stages. **a**, Shlush *et al.*[2] found that, when AML presents in the clinic with no warning, the cancer-initiating mutation in the HSC is often in the gene *DNMT3a*. Such precancerous cells are not clinically detectable. Further mutations, for example in the gene *NPM1*, then lead to AML. **b**, AML can also arise as a secondary event to myeloid-cell dysplasias, in which mature myeloid cells are not generated effectively. Subtypes of myeloid-cell dysplasias can arise from different subsets of mutations. **c**, Alternatively, further mutations in cells of chronic myeloid leukaemia (which already carry a mutation in *BCR–ABL1* or other genes) can result in more aggressive, acute-phase disease.

equally high in all but two patients studied. This suggests that both mutations were probably present in a 'founder' clone population, from which the leukaemic population expanded. From this starting point, the authors astutely deduced that the first cell to acquire a leukaemic 'driver' mutation in *DNMT3a* in these patients was a haematopoietic stem cell (HSC) — the precursors of the lymphoid and myeloid lineages — and that descendants of this cell persisted as an expanded, competitive clonal population (Fig. 1). However, whether mutation of *DNMT3a* is by itself sufficient to both initiate and sustain the growth of a clonal population before the onset of leukaemia is uncertain, because Shlush and colleagues analysed only specific genes, and so the mutational complexity of the cancers was probably underestimated[3]. Although mutation of *NPM1* can initiate AML in mouse models of the disease[4], and is often seen in humans with AML, it seems to be a secondary alteration in most cases of AML that carry a *DNMT3a* mutation.

These data agree with the results of functional experiments[5] indicating that DNMT3a normally promotes the differentiation of HSCs into other cell lineages at the expense of the HSCs' self-renewal. The mutation in *DNMT3a* in AML results in a loss of the enzyme's catalytic activity[6]. But the mutated protein can repress the function of the normal protein encoded by the non-mutated copy of *DNMT3a*, and this is expected to increase self-renewal[7]. Shlush *et al.* found mutant *DNMT3a* in a relatively high percentage of HSCs from each AML patient (up to 30%), suggesting a high degree of competitive self-renewal of the *DNMT3a*-mutant cells. However, they also observed some normal differentiation of the *DNMT3a*-mutant HSCs into lymphoid and myeloid cell lineages.

The authors confirmed this functionally by repopulating immune-cell-deficient mice with blood cells from two patients with AML. In around 75% of these mice, multiple cell lineages were established. Ten of 12 such mice were studied further, and were found to have a high frequency of the *DNMT3a* mutation but to have normal *NPM1*, indicating a functional dominance of the *DNMT3a*-mutated HSC population. Only a minority of the mice had a myeloid leukaemic population with mutations in both *DNMT3a* and *NPM1*.

Shlush and co-workers further showed that *DNMT3a*-mutant HSCs and their differentiated progeny persisted in patients' blood even when AML was in remission following chemotherapy, indicating that at least some of these pre-leukaemic ancestral cells were resistant to treatment. This may be because a high fraction is quiescent — as in other types of leukaemia and cancer. Collectively, the data highlight the persistence of benign pre-leukaemic clones. Other data point to a similar scenario, including the observation[8] of another mutation often found in AML, the *ETO–RUNX1* fusion, in HSCs and B cells. A further study[9] reported the identification of ostensibly normal HSCs, myeloid and lymphoid cells that had a subset of the mutations present in the immature AML myeloid cells of the same patient. Interestingly, and in contrast to Shlush and colleagues' findings, that study provided evidence of a linear, or evolutionary, order of several mutations in the HSC-derived pre-leukaemic cells.

More generally, these studies[2,8,9] contribute to an emerging portrait of the complexity of clonal evolutionary pathways that lead to AML[3,10,11] (Fig. 1). The premalignant phase, arising predominantly in HSCs, can be highly variable, both clinically and biologically, depending on the mutations, or combination of mutations, involved[11] and their functional impact. Common secondary mutations, such as those in *NPM1*, may result in more stringent arrest of differentiation, or more vigorous cell proliferation — steps that lead to acute-phase disease. Secondary mutations, in contrast to the founder (or very early) alterations, often seem to arise in more-lineage-committed myeloid progenitors, as demonstrated for *NPM1* by Shlush *et al.*, and previously in chronic myeloid leukaemia[12]. A study[13] of acute lymphoblastic leukaemia has similarly suggested that secondary mutations arise in more-differentiated progeny cells, and this might be expected to be a feature of other types of cancer.

As Shlush and colleagues point out, their data have several clinical implications. One is that, as a persistent and likely founder, the *DNMT3a* mutation could be both a therapeutic target and a marker for tracking residual disease. By contrast, therapies targeting secondary mutations, such as *NPM1*, will probably have only transient success. Another possibility suggested by the persistence of the pre-leukaemic clones and of mutant HSCs in remission is that these cells provide a cellular reservoir for relapse. The observation that some patients who present with *DNMT3a–NPM1* double-mutant AML but, following treatment, relapse with cancers that retain only the *DNMT3a* mutation, is compatible with this idea[14]. Finally, it would be of interest to determine how often *DNMT3a*-mutant clones arise from HSCs in ageing adults and, where such clones emerge, how frequently the evolutionary transition to AML occurs, and over what time frame. ■

Nicola E. Potter *and* **Mel Greaves** *are in the Centre for Evolution and Cancer, The Institute of Cancer Research, London SM2 5NG, UK.*
e-mails: mel.greaves@icr.ac.uk; nicola.potter@icr.ac.uk

1. Greaves, M. & Maley, C. C. *Nature* **481,** 306–313 (2012).
2. Shlush, L. I. *et al. Nature* **506,** 328–333 (2014).
3. Welch, J. S. *et al. Cell* **150,** 264–278 (2012).
4. Vassiliou, G. S. *et al. Nature Genet.* **43,** 470–475 (2011).
5. Challen, G. A. *et al. Nature Genet.* **44,** 23–31 (2012).
6. Ley, T. J. *et al. N. Engl. J. Med.* **363,** 2424–2433 (2010).
7. Kim, S. J. *et al. Blood* **122,** 4086–4089 (2013).
8. Miyamoto, T., Weissman, I. L. & Akashi, L. *Proc. Natl Acad. Sci. USA* **97,** 7521–7526 (2000).
9. Jan, M. *et al. Sci. Transl. Med.* **4,** 149ra118 (2012).
10. Walter, M. J. *et al. N. Engl. J. Med.* **366,** 1090–1098 (2012).
11. Papaemmanuil, E. *et al. Blood* **122,** 3616–3627 (2013).
12. Jamieson, C .H. M. *et al. N. Engl. J. Med.* **351,** 657–667 (2004).
13. Hong, D. *et al. Science* **319,** 336–339 (2008).
14. Krönke, J. *et al. Blood* **122,** 100–108 (2013).

This article was published online on 12 February 2014.

50 Years Ago

Discovering the Universe by Sir Bernard and Lady Lovell — Seldom has a technical exploit aroused such widespread interest and excitement as did the launching of the first Russian *Sputnik* in 1957. It was at this time that the radio telescope at Jodrell Bank was first coming into operation … So it was natural that the quickened scientific awareness of the general public should find a focal point in Prof. Lovell and the radio telescope … In view of the joint authorship one might, perhaps, anticipate some comments of a more personal nature. What does it feel like to live under the shadow of the bowl, to partner a man whose name has become a household word? Such elements are difficult to detect save in the preface, "The life which we knew before October 1957 has never quite returned". One can well imagine that this is something of an understatement.
From *Nature* 22 February 1964

100 Years Ago

'The wearing of birds' plumage — a woman's protest' — The dealers in feathers seem to think that because they have embarked on that particular trade it must never be abolished, no matter if the most exquisite birds become extinct. It is known that many trades have suffered severely from the advent of the motor-car. Whip-makers scarcely have anything to do. Harness-makers have also suffered, yet these trades could scarcely demand that motors should not be used because such might suffer thereby … There is such an abundance of lovely ornaments to be had … and if there must be feathers, then take some which require no cruelty to procure … Of course, imitation feathers would be cheap — to some women an unpardonable fault. Well, when the adornment *must* be expensive, there are jewels and laces.
From *Nature* 19 February 1914

PLASMA PHYSICS

A promising advance in nuclear fusion

Experiments conducted at the US National Ignition Facility have cleared a hurdle on the road to nuclear fusion in the laboratory, encouraging fusion scientists around the world. **SEE LETTER P.343**

MARK HERRMANN

Formidable challenges face the decades-long quest to achieve nuclear fusion — the power source of stars — in the laboratory. For a plasma to undergo self-heating nuclear fusion (ignition), it must be both hot and well confined. The facilities that hope to accomplish this goal are technological marvels, but are dauntingly expensive to build and operate. Setbacks abound as nature resists human attempts to control it, and technical hurdles lead to cost increases that threaten funding support. Thus, it is not surprising that fusion scientists throughout the world are cheering the exciting advances reported by Hurricane *et al.*[1] on page 343 of this issue. For the first time, their laboratory fusion experiment achieved more energy from fusion reactions than had been invested in the fusion fuel. These results are still a long way from ignition, but they represent a significant step forward in fusion research.

In the most commonly used laboratory fusion reaction, the fusion fuel consists of a plasma of deuterium and tritium, and the reaction produces both an α-particle (two protons and two neutrons bound together) and a neutron (Fig. 1). The charged α-particle deposits its energy locally, whereas the neutron escapes. The local heating that results from the α-particles provides the potential for ignition if energy losses from the plasma can be overcome. For a deuterium–tritium plasma at a temperature of 50 million kelvin, the criterion for achieving ignition is known as the Lawson criterion[2,3], and can be simply written as $P\tau > 25$ atm s, where P is the pressure of the plasma; τ is the energy confinement time of the plasma (the timescale over which the plasma loses its energy); atm is atmospheres; and s is seconds. Once ignition is obtained, fusion proceeds rapidly and produces much more energy than was invested in creating the plasma.

Many approaches to achieving this criterion have been conceived, but the two main methods occupy very different regimes of pressure and energy confinement time. In magnetic confinement fusion, in which a strong magnetic field confines the hot plasma, the pressure is of the order of atmospheres and the energy confinement time of the order of seconds. In inertial confinement fusion — the approach taken by Hurricane and colleagues — it is the inertia of the deuterium–tritium plasma itself that provides the confinement. In this regime, the pressure is a few times 10^{11} atm, and the energy confinement time is about 10^{-10} s.

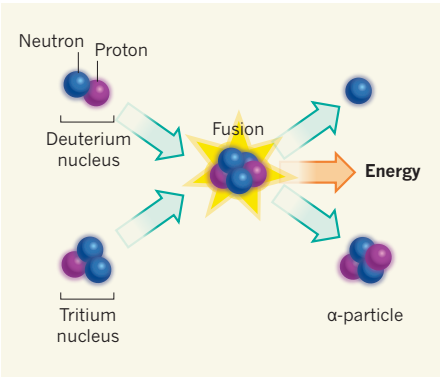


Figure 1 | Deuterium–tritium fusion reaction. When a deuterium nucleus and a tritium nucleus fuse, a neutron and an α-particle emerge and substantial energy is released. Hurricane and colleagues[1] have created a plasma of deuterium and tritium that has a fusion-energy yield greater than the energy invested in the fusion fuel.

Achieving pressures this large, even for vanishingly short times, is no easy task. Hurricane *et al.* performed their experiments using the National Ignition Facility (NIF) at the Lawrence Livermore National Laboratory in California, which was completed in 2009 after more than a decade of construction. The NIF consists of 192 laser beams that can be focused onto a centimetre-scale target containing a capsule filled with fusion fuel. The beams are capable of delivering more than 1.8 million joules of energy to the target in a carefully controlled laser pulse lasting less than 2×10^{-8} s. As extreme as this sounds, the pressure on the target is still 1,000 times lower than that needed to meet the Lawson criterion. To achieve ignition, one must compress the fusion fuel in a nearly perfect spherical implosion at velocities of hundreds of kilometres per second, while adding as little heat as possible to the fusion fuel and avoiding numerous hydrodynamic instabilities.

Hurricane *et al.* made their advance by studying deuterium–tritium implosions that were more stable than those previously explored. This increased stability was accomplished

by turning up the laser power early in time (a 'high foot') relative to previous experiments. By choosing this path, which heats the fusion fuel during the implosion, the authors gave up on the potential of achieving ignition with these implosions in the near term. However, the strategy has paid off handsomely. The fusion-energy yield at the NIF has been increased tenfold in a steady progression of experiments. The best high-foot experiment produced 17 kJ of fusion yield, which is greater than the energy invested in the fusion fuel during the implosion, and has $P\tau$ larger than 50% of what is needed for ignition. The experiments are also in much better agreement with detailed simulations, obtaining fusion-energy yields greater than 60% of what is calculated. Perhaps the most exciting observation is that calculations also suggest that the α-particle heating in these experiments is beginning to contribute significantly to further fusion-energy yield — the first hints of the self-heating process that is crucial for ignition.

Hurricane and colleagues' work comes in the wake of the National Ignition Campaign (NIC), a focused effort at the NIF from 2009 to 2012 with the goal of achieving ignition shortly after construction of the NIF was completed. Although much was learned during that period, the NIC effort fell far short of the ignition goal, and the results obtained differed considerably from predictions. The unsuccessful end of the NIC prompted congressional scrutiny of the national inertial confinement fusion programme, resulting in a substantial restructuring[4]. The restructured programme has emphasized more-stable implosions, like those pursued by Hurricane and co-workers, and more experiments dedicated to fundamental scientific issues on the path to ignition. It also underlines alternative approaches to inertial confinement fusion (currently being studied at the Omega Laser Facility in New York and at the Z Pulsed Power Facility in New Mexico) and recommends a review of progress towards ignition in 2015. These changes are paying dividends, but the 2015 review is right around the corner.

To be clear, much work remains to be done to achieve ignition. It is still not well understood why the earlier implosions studied in the NIC are so far from predictions. Issues such as the coupling of the laser energy to the target and the detailed symmetry of the implosion have been determined largely empirically. How far the high-foot implosions can be pushed remains an open question, and the goal of ignition will require a nearly 100-fold increase in fusion yield over these results. Perhaps the biggest question is: will ignition be achievable at the NIF? The answer is uncertain. Funding agencies and scientists want to know, and only time and more experiments will tell whether that will be possible. This is frequently where we who work in fusion find ourselves, facing significant obstacles but encouraged by scientific advances such as those of Hurricane and colleagues. ■

Mark Herrmann *is at the Pulsed Power Sciences Center, Sandia National Laboratories, Albuquerque, New Mexico 87185, USA.*
e-mail: mherrma@sandia.gov

1. Hurricane, O. A. *et al. Nature* **506,** 343–348 (2014).
2. Lawson, J. D. *Proc. Phys. Soc. B* **70,** 6 (1957).
3. Betti, R. *et al. Phys. Plasmas* **17,** 058102 (2010).
4. Brumfiel, G. *Nature* http://dx.doi.org/10.1038/nature.2012.12016 (2012).

This article was published online on 12 February 2014.

MOLECULAR BIOLOGY

Protein binding cannot subdue a lively RNA

Ribosomes, the cell's protein–synthesis machines, are assembled from their components in a defined order. It emerges that the first assembly step must overcome dynamic structural rearrangements. SEE ARTICLE P.334

KATHLEEN B. HALL

In cells, RNA molecules typically associate with proteins to form ribonucleoproteins. The ribosome is the large ribonucleoprotein responsible for protein synthesis, and is assembled by the sequential binding of its constituent proteins to ribosomal RNA. In this issue, Kim *et al.*[1] (page 334) describe the events that occur as the ribosomal protein S4 binds to a fragment of rRNA. The resulting complex is the first to form in the assembly of the small ribosomal subunit, and it must guide the assembly of the rest of the subunit. The authors show that the complex is an extremely dynamic structure, rather than a static building block.

Sequential association of different proteins with an RNA molecule could be a mechanism enabling RNA to adopt its functional conformation, with each protein acting as a molecular 'chaperone' to direct the folding pathway or remodel misfolded structures. As first reported[2] in 1974, the small ribosomal subunit (30S) of the bacterium *Escherichia coli* can be reconstituted *in vitro* by the sequential addition of 21 small ribosomal proteins to the 16S rRNA (1,541 nucleotides in *E. coli*). S4 is the first protein to bind, and it triggers a conformational change in the RNA that facilitates subsequent protein association.

Kim and colleagues' study is the first single-molecule analysis of S4 binding to 16S rRNA, and reveals the details of the interactions involved. Using fluorescence-based experiments, the authors watched in real time as an S4 protein bound a five-way junction in 16S rRNA, in which five RNA arms (duplexes) radiated out from a single node (the junction). They observed that the five-way junction is not a static structure passively waiting for S4 to bind and remodel it (Fig. 1). By attaching fluorescent probes to two of its arms, Kim *et al.* saw those arms move on widely different timescales. In one experiment, the arms were

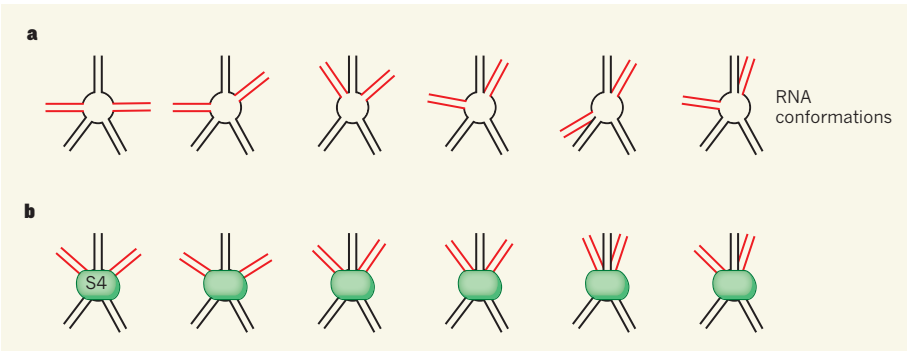


Figure 1 | Binding of the S4 protein to an RNA five-way junction. **a**, Kim *et al.*[1] studied a fragment of 16S ribosomal RNA, which consists of five duplex 'arms' radiating from a central junction, by attaching fluorescent probes to two of the arms. They observed that the RNA's structure is highly dynamic, with the labelled arms (red) adopting a wide array of conformations. The other arms are also likely to move, but were not studied. **b**, The S4 protein forms a complex with 16S rRNA in the first step of ribosome assembly by binding to its preferred RNA conformation (binding step not shown). Once bound, S4 restrains, but does not fix, the movements of the RNA arms.

closed for 100 seconds, then rapidly fluctuated between open and closed states for 50 s, before closing again for another 100 s. The authors also observed the arms jump from a closed state lasting 30 s to an open state that persisted for 15 s.

Kim and colleagues discovered a further level of complexity when they added magnesium ions (Mg^{2+}) to the system. In solutions containing concentrations of Mg^{2+} ions close to physiological levels, most of the RNA molecules existed in an open state, with the two arms far apart. But at a higher concentration, two populations of RNA existed, one in an open state and the other, slightly larger population in a closed state; the RNA molecules alternated stochastically between these states.

S4 specifically recognizes the junction and contacts an arm[3], but if the RNA structure changes on millisecond and second timescales, as seen in the authors' real-time fluorescence data, how can the protein find its binding site? Proteins that bind to specific sequences in RNAs do so by recognizing single-stranded regions of the molecules. Such regions are intrinsically flexible, and the difference between their free and protein-bound structures is often dramatic. The mechanism of RNA binding by a protein must therefore include a means of catching a conformation that displays the RNA nucleotides in a geometry that the protein can recognize.

One widely used model of how a protein might bind to a flexible RNA is called conformation capture[4–6]. This model acknowledges that RNA in solution is best described as an ensemble of conformations that have an unknown population distribution. Assuming that the structures are nearly equal in energy, and that they interconvert, then only some of the molecules will be able to bind to the protein. When the protein encounters an RNA with a binding-competent structure, it captures it, forming a complex. After capture, the RNA's structure changes to accommodate the protein's binding site. Such a mode of complex formation, in which the protein manipulates the RNA to complement its surface, is known as induced fit[7–9].

The conformation-capture and induced-fit models neglect the probability that the protein also undergoes conformational changes, so that in reality the capture and fitting processes are mutual. The authors show that S4–rRNA binding requires that S4 select among rRNA conformations (conformational selection) but that, when S4 is bound, new patterns of RNA dynamics appear in the complex (induced conformational changes). Such dynamics might be essential for the addition of the next protein.

Kim *et al*. describe the kinetics of the changes in the RNA, and the populations of molecules in each conformational state (in this case, two free and two bound states predominate). They also show how those kinetics and populations are altered by Mg^{2+}: S4 binds most efficiently at high Mg^{2+} concentrations, which hold the RNA in a closed conformational state. However, the authors observe that the RNA in the bound state is not a static prisoner — at least one of its arms waves frantically around the protein, as if in protest at its capture.

The authors make these experiments look easy. They are not. But they show the heart of RNA–protein interactions in a way that other methods cannot. In this case, Kim *et al*. have demonstrated that S4–rRNA mutual recognition is not a simple lock-and-key process, but a much more challenging one that could be described as conformational capture followed by mutual induced fit, and which thereby allows both protein and RNA to retain dynamic motion. S4–rRNA recognition may well be an archetype of the structure and dynamics of ribonucleoproteins. If so, these experiments illustrate what is really meant by "the protein binds the RNA". ■

Kathleen B. Hall *is in the Department of Biochemistry and Molecular Biophysics, Washington University Medical School, St Louis, Missouri 63110, USA.*
e-mail: kathleenhal@gmail.com

1. Kim, H. *et al*. *Nature* **506,** 334–338 (2014).
2. Held, W. A., Ballou, B., Mizushima, S. & Nomura, M. *J. Biol. Chem.* **249,** 3103–3111 (1974).
3. Stern, S., Wilson, R. C. & Noller, H. F. *J. Mol. Biol.* **192,** 101–110 (1986).
4. Boehr, D. D., Nussinov, R. & Wright, P. E. *Nature Chem. Biol.* **5,** 789–796 (2009).
5. Haller, A., Rieder, U., Aigner, M., Blanchard, S. C. & Micura, R. *Nature Chem. Biol.* **7,** 393–400 (2011).
6. Leulliot, N. & Varani, G. *Biochemistry* **40,** 7947–7956 (2001).
7. Koshland, D. E. Jr *Proc. Natl Acad. Sci. USA* **44,** 98–104 (1958).
8. Williamson, J. R. *Nature Struct. Mol. Biol.* **7,** 834–837 (2000).
9. Weikl, T. R. & von Deuster, C. *Proteins* **75,** 104–110 (2009).

This article was published online on 12 February 2014.

AGEING

Genetic rejuvenation of old muscle

In advanced age, the stem cells responsible for muscle regeneration switch from reversible quiescence to irreversible senescence. Targeting a driver of senescence revives muscle stem cells and restores regeneration. SEE ARTICLE P.316

MO LI & JUAN CARLOS IZPISUA BELMONTE

One of the telltale signs of advanced ageing is loss of skeletal-muscle mass and strength, a phenomenon known as sarcopenia. Muscle strength is inversely correlated with mortality in old populations[1,2], and the decline in strength is attributable to the decreased regenerative capacity of muscle stem cells, called satellite cells, with age. Whether this decline is caused by cell-intrinsic and/or environmental alterations has remained unclear. On page 316 of this issue, Sousa-Victor *et al*.[3] shed light on this debate by uncovering intrinsic aspects of age-related satellite-cell dysfunction that account for the loss of muscle maintenance (homeostasis) and regeneration. This study also provides a potential strategy for satellite-cell rejuvenation that could benefit geriatric individuals and those with progeria, a disorder in which cells age prematurely.

Satellite cells reside in close proximity to large muscle cells called myofibres. They are responsible for post-natal muscle growth, as well as regeneration after injury. Because under normal conditions skeletal muscles exhibit low turnover, satellite cells exist in a reversible quiescent state. Once stimulated by homeostatic demand or damage, they become activated and re-enter the cell cycle to generate muscle progenitor cells — which differentiate and fuse to form new fibres — or self-renew to replenish the stem-cell population (Fig. 1a).

Satellite cells actively maintain their quiescent state through transcriptional, post-transcriptional and post-translational mechanisms[4,5]. Dysregulation of quiescence often leads to stem-cell exhaustion and failure of regeneration[6]. Evidence suggests[7] that extrinsic changes in the milieu surrounding satellite cells contribute to the cells' well-documented decline during ageing, and that exposure to a youthful environment can reverse this process. Notably, fibroblast growth factor-2 produced by neighbouring cells in an aged environment disrupts satellite-cell quiescence and self-renewal[8]. So far, few studies have investigated the intrinsic changes in satellite cells during ageing[5,9] or the potential role of such changes in regenerative decline.

To address these issues, Sousa-Victor and colleagues first compared muscle properties in mice of various ages and established that features of sarcopenia appeared mainly in mice of geriatric age (at 28 months and thereafter) or in a mouse model of progeria. Notably, on the induction of muscle injury by a toxin, the authors observed a sharp decline

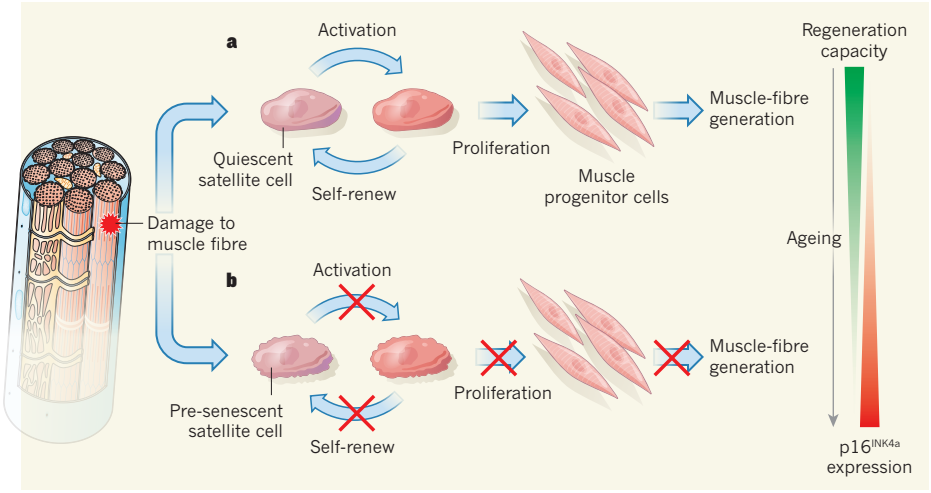


Figure 1 | Old age disrupts muscle regeneration. **a**, Satellite cells, a type of muscle stem cell, remain quiescent under normal conditions. After muscle damage, satellite cells become activated and re-enter the cell cycle to produce muscle progenitor cells that regenerate new muscle fibres. They also self-renew to replenish the stem-cell population. **b**, Sousa-Victor *et al.*[3] report that during ageing, geriatric satellite cells lose their reversible quiescent state owing to derepression of the gene encoding $p16^{INK4a}$, a regulator of cellular senescence. Instead, they adopt a senescent-like state (becoming pre-senescent cells), which impairs the regeneration process, including activation, proliferation and self-renewal.

in the regenerative capacity of satellite cells in geriatric, sarcopenic mice compared with old, non-sarcopenic mice. This phenomenon cannot be explained by a reduced satellite-cell pool, because the number of these cells was comparable in both groups of mice.

Next, the authors conducted a series of experiments in which satellite cells from geriatric and old animals were transplanted into young mice, and this definitively proved that the regenerative decline of geriatric muscle is due to changes intrinsic to satellite cells, independent of the host environment. Intriguingly, geriatric satellite cells exhibited a cell-cycle block and defective activation in response to injury both *in situ* and after transplantation, indicating a failure to maintain a reversible state of quiescence.

What factors could be responsible for this loss of quiescence? Through comparative analyses of the gene-expression programs of quiescent satellite cells of different ages, Sousa-Victor and co-workers narrowed down the list of candidates to the tumour-suppressor protein $p16^{INK4a}$, which is regarded as a master regulator of cellular senescence. In a series of experiments, the authors found evidence to support a link between *p16*^{INK4a} derepression and defective satellite-cell activation.

In a mouse model that underwent successive rounds of injury, the authors observed a depletion of self-renewing geriatric satellite cells over time, whereas normal satellite cells continued to self-renew. The pressure to proliferate in response to injury drove geriatric satellite cells into full-blown senescence, as evidenced by the expression of several classic markers of senescence. This correlated with reduced levels of phosphorylated retinoblastoma (Rb) protein, and with reduced expression of genes regulated by Rb and the transcription factor E2F, suggesting that the well-defined $p16^{INK4a}$/Rb/E2F signalling axis drives the conversion to senescence.

Sousa-Victor *et al.* genetically silenced $p16^{INK4a}$ expression and found that this restored self-renewal and proliferation in geriatric satellite cells. These results show that $p16^{INK4a}$ derepression in geriatric and progeric satellite cells leads to the loss of the reversible quiescent state and to the adoption of a senescent-like state, which impairs regeneration (Fig. 1b). The relevance of this work to human health is strengthened by Sousa-Victor and co-workers' finding that the $p16^{INK4a}$/Rb/E2F axis drives dysfunction in geriatric human satellite cells similarly to the way it does in mice.

Although $p16^{INK4a}$ expression during ageing has been shown to impair regeneration in blood, neural and pancreatic tissues[6], it has never been reported in aged satellite cells, despite previous gene-profiling studies[4]. The use of a clearly defined sarcopenic geriatric population may be the key to this discovery, which itself represents an important addition to a growing body of evidence[10,11] showing that $p16^{INK4a}$-induced senescence limits the regenerative capacity of stem cells during ageing and contributes to age-related pathologies. Because $p16^{INK4a}$ expression is also a barrier to stem-cell reprogramming[12,13], this research increases the potential benefits of transiently inactivating $p16^{INK4a}$ for regenerative medicine.

Sousa-Victor and colleagues' study provides a new view of satellite-cell ageing, but the results inevitably raise further questions. For example, what triggers the $p16^{INK4a}$/Rb/E2F senescence pathway during ageing? A recent study[9] found no evidence of significant accumulation of DNA damage in old satellite cells compared with young ones. Could it be that $p16^{INK4a}$ is derepressed owing to signals from neighbouring senescent cells, such as low-level systemic inflammation or elevated levels of reactive oxygen species?

Because satellite cells are not a uniform population, it is possible that a sub-population is more susceptible or immune to the quiescent-to-senescent switch. Along this line, it will be interesting to determine whether geriatric satellite cells that are activated on injury maintain full 'stemness'. Could any as-yet-unidentified, age-associated environmental factors be neutralized to postpone the $p16^{INK4a}$ induction in satellite cells of sarcopenic muscle? And, if so, could physical exercise delay $p16^{INK4a}$ induction?

Finally, this study presents yet another addition to the list of potential strategies to improve the regenerative capacity of aged tissue[11,14,15]. It may be worth considering whether the benefits of transiently reducing tumour-suppressor levels in stem cells outweigh the associated risks, in the context of preventing an age-related decline in regenerative potential. Whether these strategies can be safely implemented in the clinic to maximize human health span deserves thorough investigation in the near future. ■

Mo Li *and* **Juan Carlos Izpisua Belmonte**
are in the Gene Expression Laboratory, Salk Institute for Biological Studies, La Jolla, California 92037, USA.
e-mail: belmonte@salk.edu

1. Ruiz, J. R. *et al. Br. Med. J.* **337,** a439 (2008).
2. de Brito, L. B. B. *et al. Eur. J. Prev. Cardiol.* http://dx.doi.org/10.1177/2047487312471759 (2012).
3. Sousa-Victor, P. *et al. Nature* **506,** 316–321 (2014).
4. Cheung, T. H. & Rando, T. A. *Nature Rev. Mol. Cell Biol.* **14,** 329–340 (2013).
5. Liu, L. *et al. Cell Rep.* **4,** 189–204 (2013).
6. Orford, K. W. & Scaden, D. T. *Nature Rev. Genet.* **9,** 115–128 (2008).
7. Jang, Y. C., Sinha, M., Cerletti, M., Dall'Osso, C. & Wagers, A. J. *Cold Spring Harb. Symp. Quant. Biol.* **76,** 101–111 (2011).
8. Chakkalakal, J. V., Jones, K. M., Basson, M. A. & Brack, A. S. *Nature* **490,** 355–360 (2012).
9. Cousin, W. *et al. PLoS ONE* **8,** e63528 (2013).
10. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. *Cell* **153,** 1194–1217 (2013).
11. Baker, D. J. *et al. Nature* **479,** 232–236 (2011).
12. Li, H. *et al. Nature* **460,** 1136–1139 (2009).
13. Menendez, S. *et al. Aging Cell* **11,** 41–50 (2012).
14. Pajcini, K. V., Corbel, S. Y., Sage, J., Pomerantz, J. H. & Blau, H. M. *Cell Stem Cell* **7,** 198–213 (2010).
15. Shyh-Chang, N. *et al. Cell* **155,** 778–792 (2013).

This article was published online on 12 February 2014.

CORRECTION

In the News & Views article 'Conservation: Making marine protected areas work' by Benjamin S. Halpern (*Nature* **506,** 167–168; 2014), Figure 1 was published with the wrong caption. The correct caption can be seen in the online version at go.nature.com/pssric.

REVIEW

doi:10.1038/nature13068

The rise of oxygen in Earth's early ocean and atmosphere

Timothy W. Lyons[1], Christopher T. Reinhard[1,2,3] & Noah J. Planavsky[1,4]

The rapid increase of carbon dioxide concentration in Earth's modern atmosphere is a matter of major concern. But for the atmosphere of roughly two-and-half billion years ago, interest centres on a different gas: free oxygen (O_2) spawned by early biological production. The initial increase of O_2 in the atmosphere, its delayed build-up in the ocean, its increase to near-modern levels in the sea and air two billion years later, and its cause-and-effect relationship with life are among the most compelling stories in Earth's history.

Most of us take our richly oxygenated world for granted and expect to find O_2 everywhere—after all, it makes up 21% of the modern atmosphere. But free oxygen, at levels mostly less than 0.001% of those present in the atmosphere today, was anything but plentiful during the first half of Earth's 4.5-billion-year history. Evidence for a permanent rise to appreciable concentrations of O_2 in the atmosphere some time between 2.4 and 2.1 billion years (Gyr) ago (Fig. 1) began to accumulate as early as the 1960s[1]. This step increase, now popularly known as the 'Great Oxidation Event' or GOE[2,3], left clear fingerprints in the rock record. For example, the first appearance of rusty red soils on land and the disappearance of easily oxidized minerals such as pyrite (FeS_2) from ancient stream beds[3,4] both point to the presence of oxygen in the atmosphere. The notion of a GOE is now deeply entrenched in our understanding of the early Earth, with only a few researchers suggesting otherwise[5].

Far more controversial is the timing of the first emergence of O_2-producing photosynthesis, the source of essentially all oxygen in the atmosphere. Among the key questions is whether this innovation came before, or was coincident with, the GOE. Tantalizing organic geochemical data pinpointed pre-GOE O_2 production[6], but subsequent claims of contamination cast doubt[7,8]. Recently, new inorganic approaches have restored some of that lost confidence[9], and assertions of pre-GOE oxygenesis have bolstered research[10,11] that explores buffers or sinks, whereby biological O_2 production was simultaneously offset by consumption during reactions with reduced compounds emanating from Earth's interior (such as reduced forms of hydrogen, carbon, sulphur and iron). Delivery of these oxygen-loving gases and ions to the ocean and atmosphere, tied perhaps to early patterns of volcanism and their relationships to initial formation and stabilization of the continents[10,11], must have decreased through time to the point of becoming subordinate to O_2 production, which may have been increasing at the same time. This critical shift triggered the GOE. In other words, buffering reactions that consumed O_2 balanced its production initially, thus delaying the persistent accumulation of that gas in the atmosphere. Ultimately, however, this source–sink balance shifted in favour of O_2 accumulation—probably against a backdrop of progressive loss of hydrogen (H_2) to space, which contributed to the oxidation of Earth's surface[12–14]. Other researchers have issued a minority report challenging the need for buffers, arguing instead that the first O_2-yielding photosynthesis was coincident with the GOE[15].

As debate raged over the mechanistic underpinnings of the GOE, there emerged a far less contentious proof (a 'smoking gun') of its timing—namely, the disappearance of distinctive non-mass-dependent (NMD) sulphur isotope fractionations in sedimentary rocks deposited after about 2.4–2.3 Gyr ago[16] (Fig. 2). Almost all fractionations among isotopes of a given element scale to differences in their masses; NMD fractionations deviate from this typical behaviour. The remarkable NMD signals are tied to photochemical reactions at short wavelengths involving gaseous sulphur compounds released from volcanoes into the atmosphere. For the signals to be generated and then preserved in the rock record requires extremely low atmospheric oxygen levels, probably less than 0.001% of the present atmospheric level (PAL)[17], although other properties of the early biosphere, such as atmospheric methane abundance[18,19] and biological sulphur cycling[20], certainly modulated the NMD signal.

Aware of the possibility that the 'Great' in GOE may exaggerate the ultimate size of the O_2 increase and its impact on the ocean, Canfield[21] defined a generation of research by championing the idea that ultimate oxygenation in the deep ocean lagged behind the atmosphere by almost two billion years. Finding palaeo-barometers for the amount (or partial pressure) of O_2 in the ancient atmosphere is a famously difficult challenge, but the implication is that oxygen in the atmosphere also remained well below modern levels (Fig. 1) until it rose to something like modern values about 600 million years (Myr) ago. In this view, this second O_2 influx oxygenated much of the deep ocean while enriching the surface waters, thus welcoming the first animals and, soon after, their large sizes and complex ecologies above and within the sea floor.

From this foundation, a fundamentally new and increasingly unified model for the rise of oxygen through time is coming into focus (Fig. 1). Our story begins with the timing of the earliest photosynthetic production of oxygen and its relationship to the sulphur isotope record. After the GOE, we assert that oxygen rose again and then fell in the atmosphere and remained, with relatively minor exceptions, at extremely low levels for more than a billion years. This prolonged stasis was probably due to a combination of fascinating biogeochemical feedbacks, and those conditions spawned an oxygen-lean deep ocean. This anoxic ocean probably harboured sufficiently large pockets of hydrogen sulphide to draw down the concentrations of bioessential elements and thus, along with the overall low oxygen availability, challenge the emergence and diversification of eukaryotic organisms and animals until the final big step in the history of oxygenation and the expansion of life. All of this evidence comes from very old rocks, which present unique challenges—not the least of which is that constant recycling at and below Earth's surface erases most of the record we seek. But with challenge comes opportunity.

[1]Department of Earth Sciences, University of California, Riverside, California 92521, USA. [2]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA. [3]School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, Georgia 30332, USA. [4]Department of Geology and Geophysics, Yale University, New Haven, Connecticut 06511, USA.

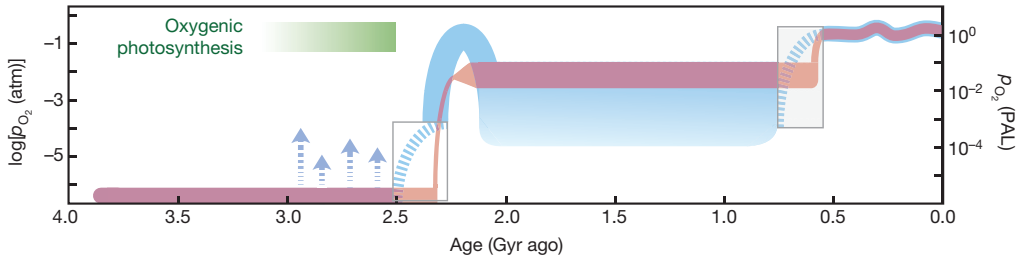


Figure 1 | Evolution of Earth's atmospheric oxygen content through time. The faded red curve shows a 'classical, two-step' view of atmospheric evolution[95], while the blue curve shows the emerging model (p_{O_2}, atmospheric partial pressure of O_2). Right axis, p_{O_2} relative to the present atmospheric level (PAL); left axis, log p_{O_2}. Arrows denote possible 'whiffs' of O_2 late in the Archaean; their duration and magnitude are poorly understood. An additional frontier lies in reconstructing the detailed fabric of 'state changes' in atmospheric p_{O_2}, such as occurred at the transitions from the late part of the Archaean to the early Proterozoic and from the late Proterozoic to the early Phanerozoic (blue boxes). Values for the Phanerozoic are taken from refs 96 and 97.

The first oxygen from photosynthesis

Because oxygenic photosynthesis is the only significant source of free oxygen on Earth's surface, any evaluation of our planet's oxygenation history must begin by asking when this metabolism evolved. Yet despite decades of intensive investigation, there is no consensus. Current estimates span well over a billion years—from ~3.8 (ref. 22) to 2.35 (ref. 15) Gyr ago—almost one-third of Earth's history. Part of the problem lies with difficulties in differentiating between oxidation pathways that can be either biotic or abiotic and can occur with and without free oxygen. Banded iron formations, for example, are loaded with iron oxide minerals that often give these ancient deposits their spectacular red colours. The prevailing view for many years was that microbial oxygen production in the shallow ocean was responsible for oxidizing iron, which was locally abundant in the otherwise oxygen-free ocean. More recent studies, however, explain this iron oxidation without free O_2—specifically, through oxidation pathways requiring only sunlight (ultraviolet oxidation[23] and anoxygenic photosynthesis[24,25]). Microbial fossils of Archaean age (older than 2.5 Gyr; see Fig. 2 for time units) have very simple morphologies, and it is therefore difficult to link them to specific metabolisms, such as oxygen-producing photosynthesis. Similarly, the significance, and even the biogenicity, of Archaean stromatolites and microbially induced sedimentary structures have long been debated[26].

Other researchers vied to find more definitive indicators of microbial oxygen production. Among them, Brocks *et al.*[6] published organic biomarker data thought to record the presence of cyanobacteria and eukaryotes in 2.7-Gyr-old rocks. Biomarkers are molecular fossils derived from primary organic compounds that, in the best case, can be tied uniquely to specific biological producers present at the time the sediments were deposited. Cyanobacteria are important because they were the earliest important producers of O_2 by photosynthesis. Recognition of sterane biomarkers from eukaryotes strengthens the identification of oxygen production because O_2 is required, albeit at very low levels[27], for biological synthesis of their sterol precursors. If correct, these data would extend the first production and local accumulation of oxygen in the ocean to almost 300 Myr before the GOE as it is now popularly defined (that is, based on the disappearance of NMD fractionations of sulphur isotopes). Contrary studies, however, argue that O_2 is not required to explain these particular biomarkers[15]; others challenge the integrity of the primary signals, suggesting later contamination instead[8]. Very recent results from ultraclean sampling and analysis also raise serious concern about the robustness of the biomarker record during the Archaean[28]—and in particular point to contamination for the results of Brocks *et al.*[6] Ironically, some the best earliest organic evidence for oxygenic photosynthesis may lie more with the common occurrence of highly organic-rich shales of Archaean age than with sophisticated biomarker geochemistry (Box 1).

Over the past decade, a body of trace-metal and sulphur data has grown—independent of the biomarker controversy—that also points to oxygen production long before the disappearance of NMD sulphur isotope fractionations (Fig. 2). This evidence for early oxygenesis allows for at least transient accumulation of the gas in the atmosphere and even for hotspots of production in local, shallow, cyanobacteria-rich marine oases[29]. Despite some controversy surrounding these inorganic proxy approaches (reviewed in ref. 30), many researchers interpret strong trace-metal enrichments in marine sediments as convincing signatures of significant oxidative weathering of pyrite and other sulphide minerals on land long before the GOE—implying O_2 accumulation in the atmosphere. Sulphide minerals in the crust are often enriched in the metals of interest, such as molybdenum (Mo) and rhenium (Re), and when oxidized those metals are released to rivers and ultimately the ocean.

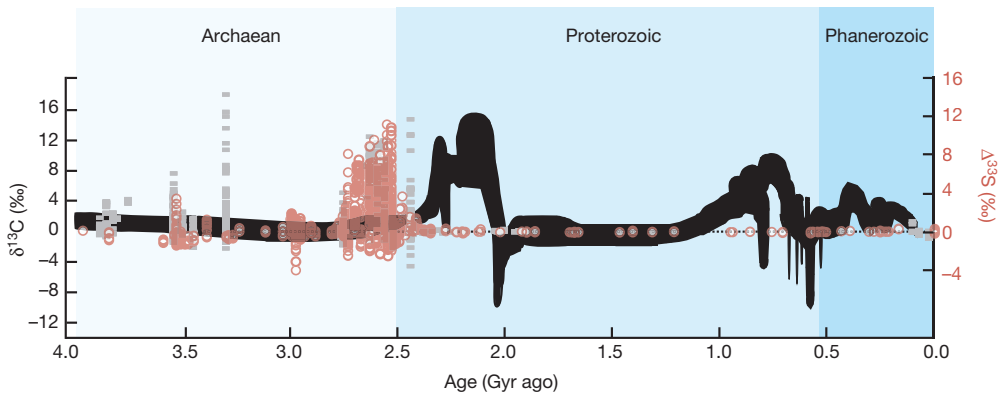


Figure 2 | Summary of carbon (black) and sulphur (red and grey) isotope data through Earth's history. Data are shown as $\delta^{13}C$ (left axis and $\Delta^{33}S$ ($= \delta^{33}S - 0.515\delta^{34}S$; right axis). Grey sulphur data were generated by secondary ion mass spectrometry (SIMS); red circles designate all other data—bulk and small sample (micro-drilled and laser) analyses. Notable features include the large range of $\Delta^{33}S$ values during Archaean time, the large $\delta^{13}C$ excursion during the early Proterozoic, relative stasis in $\delta^{13}C$ during the mid-Proterozoic, and the large negative $\delta^{13}C$ excursions during the late Proterozoic. Data are from references as compiled in refs 33 and 53.

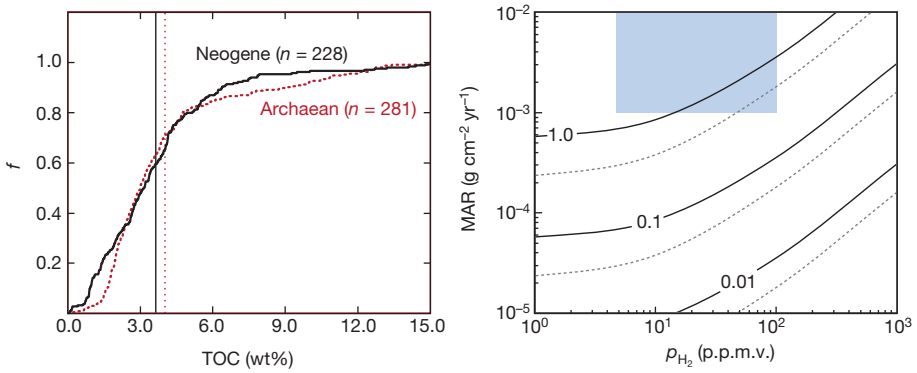
BOX 1

Evidence for oxygen-producing photosynthesis before the GOE

In the face of recent challenges to the Archaean biomarker record, the abundant organic matter from this interval takes on a more general importance. Specifically, how was this copious organic matter produced, and was O_2 a by-product? Photosynthetic life requires both light and a source of reducing power—an electron donor. Because the ubiquitous H_2O molecule is the electron donor for oxygenic photosynthesis, it is reasonable to expect that the initiation of oxygenic photosynthesis would 'supercharge' carbon fluxes through the biosphere. Nevertheless, organic-rich shales are a common component of the Archaean rock record, and the amount of total organic carbon (TOC) of pre-GOE (Archaean) shales is indistinguishable from TOC recorded in similar modern and near-modern environments (Box 1 Figure, left).

Three alternative electron donors could power delivery of significant quantities of organic carbon to marine sediments during the Archaean without releasing O_2: hydrogen sulphide (H_2S), ferrous iron (Fe^{2+}) and molecular hydrogen (H_2). Photosynthesis based on H_2S is difficult to maintain at steady state without an external carbon source[99], and many organic-rich Archaean shales were deposited from Fe^{2+}-containing waters[32], arguing against an H_2S-based pathway. Photosynthesis based on Fe^{2+} is another possibility, but this metabolism generates two physically associated particulate species (organic carbon and solid Fe-oxide minerals) at a relatively constant ratio, and these will mutually annihilate through microbial iron reduction at roughly the same ratio.

H_2-based photosynthesis is more difficult to assess. We can, however, obtain some estimates of the TOC values as a function of H_2 fluxes to the photic zone (Box 1 Figure, right). Even given the very conservative assumptions used here, it is difficult to explain typical Archaean TOC values by H_2-based photosynthesis, let alone the most elevated values from the record. We are thus left with oxygenic photosynthesis as the most likely explanation for organic-rich shales in the pre-GOE ocean.



Box 1 Figure | The significance of organic carbon content in sedimentary rocks of Archaean age (>2.5 Gyr old). Left, cumulative frequency (*f*) distributions for the total organic carbon content (TOC) of Neogene/recent (black trace) and Archaean (dotted red trace) organic-rich sedimentary rocks from references as compiled in ref. 60. Also shown are the overall average TOC contents for the two data sets (vertical lines). Note that the data for the two time periods are virtually identical. Right, the combinations of atmospheric H_2 content (partial pressure of H_2, p_{H_2}) and sediment mass accumulation rate (MAR) required to attain a given TOC value (TOC = $flux_{H2}$/MAR). Black solid contours correspond to a TOC value of 5 wt%, while grey dashed contours correspond to a value of 10 wt%. The shaded blue box denotes a plausible range for these two parameters, assuming a shelf-to-outer slope depositional setting[100] and results from Archaean ecosystem modelling[101]. Contours are labelled according to the preservation efficiency of organic carbon (that is, a value of 1.0 refers to 100% preservation). For comparison, the preservation efficiency of carbon produced in surface waters in modern anoxic basins (that is, where preservation efficiency is highest in the modern ocean) are of the order of ~1–2% (ref. 102). We assume a vertical advection rate of 1.0 m d^{-1}, typical of regions of vigorous upwelling in the modern coastal ocean[103], and an elevated deep-ocean H_2 concentration of 100 nM, both of which are extremely conservative for our purposes.

The most publicized examples of such diagnostic metal enrichments—the so-called whiffs of oxygen—come from 2.5-Gyr-old organic-rich shales drilled in Western Australia. All Archaean rocks have experienced complex histories at and beneath Earth's surface, and it is important to consider the potential overprints on primary geochemical records during and after burial[9]. However, no coherent secondary alteration model has yet emerged to explain the 'whiff' metal enrichment patterns, particularly given their strikingly sympathetic behaviour with other, independent indicators of depositional chemistry and the rhenium–osmium systematics that yield both robust depositional ages for the rocks and persuasive evidence against appreciable alteration[9,31]. Parsimony currently lies with O_2-related processes.

It may at first seem counterintuitive to suggest that O_2 was oxidizing pyrite and other sulphide minerals, which freed up trace metals for delivery to the ocean by rivers, beneath an atmosphere presumed to have had very low O_2 levels—perhaps much less than 0.001% of PAL. However, such oxidation is possible with only subtle increases in atmospheric O_2 content[9,32]. Also, recent results allow for another intriguing possibility: once NMD signals that formed in an oxygen-poor atmosphere were captured in pyrite and other minerals in sedimentary rocks, they would have been recycled when those rocks were later uplifted as mountain ranges and the pyrite was oxidized[33]. In other words, rivers may have delivered recycled sulphur with a strong NMD signal to the ocean, which can be captured in coeval sediments, long after O_2 rose, either transiently or permanently, to a point that precluded additional signal generation and preservation in the atmosphere. This 'crustal memory effect' allows for the possibility of large and persistent increases of atmospheric oxygen for tens of millions of years or more without complete loss of the NMD fingerprint; it would have taken repeated cycles of weathering, dilution, burial and uplift beneath an oxygenated atmosphere to erase the NMD signal completely. The message is that sulphur isotope records of NMD fractionation, when viewed through the filter of sedimentary recycling, may complicate efforts to date the GOE precisely, and atmospheric oxygen levels for periods of the Archaean may have been much higher than previously imagined. That said, the broad cause-and-effect relationships remain intact: more conventional mass-dependent

sulphur isotope records, which roughly track the availability of sulphate in the ocean and thus oxygen in the ocean–atmosphere system and related microbial activity without recycling artefacts, show at least general agreement with the NMD signal and dramatic and probably coupled climate change[21,34]. Further work on the early sulphur cycle will more firmly establish the isotope distributions among the various surface reservoirs and thus refine the potential importance of early recycling as an overprint on the atmospheric NMD record.

The GOE

In light of these new perspectives, the GOE might be best thought of as a protracted process rather than a discrete event marking the loss of NMD sulphur fractionations from the sedimentary record. The GOE defined this way becomes a transitional interval of yo-yo-ing biospheric oxygenation[5] during which the ups and downs of O_2 concentrations in the atmosphere reflected a dynamic balance between time-varying early oxygen production and its concurrent sinks—a scenario more consistent with Holland's initial definition of an extended GOE[2]. It is likely that the sources overcame the sinks, at first intermittently and then permanently. And any volatility in atmospheric oxygen content, reflecting perhaps trace-gas behaviour with a relatively short residence time, could be blurred in the NMD sulphur record by sedimentary recycling. Based on available evidence, this critical transitional period took place between roughly 2.5 and 2.3 Gyr ago[34–36], but suggestions of oxygenic photosynthesis much older than 2.5 Gyr ago, although not beyond dispute, are emerging[37] and challenging our conventional views of the GOE.

As stressed above, Earth's O_2 ultimately comes from photosynthesis. In the ocean today, as in the past, the lion's share of that O_2 is just as quickly consumed through decay—or more specifically, through aerobic microbial respiration. For the atmosphere to receive a boost in its oxygen content, some of that primary production in the surface ocean must escape this short-term recycling and become buried long-term beneath the sea floor. This organic-carbon burial changes the stable isotopic composition of dissolved inorganic carbon (ΣCO_2) in the ocean because the organic matter has a lower ratio of $^{13}C/^{12}C$ compared to the remaining inorganic carbon in the host sea water. This fractionation occurs during photosynthetic carbon fixation. The standard view is that the varying carbon isotope composition of sea water, recorded often with fidelity in limestone and dolostone (a magnesium-rich carbonate rock), should track temporal patterns of organic-carbon burial. For example, a dramatic increase in organic burial should manifest in a positive carbon isotope excursion. This approach has been used widely to estimate carbon burial and the O_2 content of the atmosphere through time[38]. Although the carbon isotope details of this transition are a work in progress, and emerging data are pointing to early isotope shifts[34], there is at present no evidence for a large, globally synchronous positive $\delta^{13}C$ shift in carbonate rocks across the GOE transition (Fig. 2) as defined by the permanent loss of NMD sulphur signals—suggesting that it is not a simple matter of a big increase in organic burial as the trigger.

As a corollary to the idea of O_2 production well before the GOE, a balance between carbon burial and compensatory buffering must have initially permitted appreciable oxygen production via photosynthesis without permanent accumulation in the atmosphere[10,11,13,18,39] (Fig. 1). Recent buffer models generally assume that the redox state of the mantle and magmas derived from it did not change significantly leading up to the GOE[40–42]—an idea that no doubt will be revisited in future work. From this position, these models instead emphasize decreases in delivery of reduced gases (H_2 and S species, in particular) and thus waning O_2 buffer capacity as a function of fundamental shifts in the nature of volcanoes. More to the point, a shift from dominantly submarine to increasingly subaerial volcanism as continents grew and stabilized could have led to release of more oxidized gases[10,11]. If correct, the broad temporal overlap of the GOE and first-order tectonic reorganization classically assumed to mark the Archaean–Proterozoic boundary is anything but a coincidence, and the magnitude of the NMD sulphur isotope anomaly through this transition probably varied in part with tectonic controls on volcanic release of sulphur-bearing gases[20]. Various nutrient-based buffering scenarios have also been proposed, and these too may link to long-term trends in volcanism[43]. Regardless of the specific buffer(s), and absent evidence for dramatic increases in organic burial, the balance between sources and sinks ultimately tipped in favour of photosynthetic production perhaps tens of millions of years before the permanent loss of the NMD sulphur isotope signal in rocks dating from 2.4 to 2.3 Gyr ago—and transiently perhaps hundreds of millions of years earlier.

That the first of the great 'Snowball Earth' glaciations is roughly coincident with the GOE[1,44] is probably no coincidence either. Most models for the pre-GOE atmosphere assert that comparatively large amounts of methane (CH_4), along with higher hydrocarbon gases such as ethane (C_2H_6) resulting from methane photochemistry, were produced and persisted under the generally low sulphate (SO_4^{2-}) conditions of the Archaean ocean and low O_2 in the ocean and atmosphere[45–48]. Methane is readily oxidized in the presence of free oxygen, as well as in the absence of oxygen (anaerobically) when coupled to microbial reduction of a number of different oxidants, most notably sulphate[49]. Also, in the absence or near absence of oxygen and sulphate, a greater amount of labile organic matter is available for microbial methane production (methanogenesis). Imagine a pre-GOE world, then, with mostly vanishingly small amounts of O_2 in the ocean and atmosphere; the ocean was dominated instead by high dissolved iron concentrations and the atmosphere by high methane and ethane with residence times perhaps orders of magnitude longer than today's. An important side issue here is that sulphate, which abounds in the ocean today, derives mostly from oxidation of pyrite on the continents in the presence of O_2, like the trace metals discussed earlier.

Methane and its photochemical products deserve our special attention because their roles as greenhouse gases may very well have helped to keep the early Earth habitable (by maintaining a liquid ocean) in the face of a Sun that was only about 70% to 80% as luminous as it is today[50]. This, of course, is the faint young Sun paradox discussed by Sagan[51] and many others. It follows from our understanding of the GOE that the rising O_2 content of the atmosphere might have displaced methane and other hydrocarbons, as well as H_2, as the dominant redox gas, leading to crashing temperatures and plunging the Earth into its first great 'Snowball Earth' ice age. And the timescales of atmospheric oxygenation, particularly when we consider the possibility of temporal blurring of the GOE in light of NMD sulphur recycling, may indeed mesh with the geologic record of early glaciation.

In the wake of the GOE

Until recently, the widely accepted timeline regarding O_2 was that its concentration rose in the atmosphere only modestly at the GOE and waited patiently for almost two billion years before it climbed higher (Fig. 1). Several new studies, however, are suggesting a far more dynamic screenplay, with the possibility of a much larger increase early on and then a deep plunge to lower levels that extended over a few hundred million years after the onset of the GOE (Fig. 1). These scenes play out in the most prominent positive carbon isotope event in Earth's history—the Lomagundi excursion observed around the world in rocks dating from roughly 2.3 to 2.1 Gyr ago with $\delta^{13}C$ values extending well beyond +10‰ (ref. 52; Fig. 2).

Despite earlier occurrences of markedly positive carbonate $\delta^{13}C$ values[34], the onset of the Lomagundi excursion proper appears after widespread glaciation and the loss of NMD sulphur fractionations (Fig. 2). The anomalous carbon isotope behaviour of the Lomagundi excursion is most parsimoniously tied to intense burial of organic matter[53] rather than reflecting diagenetic carbonate precipitation, as previously proposed[54]. Assuming the Lomagundi excursion is tied to organic burial, the carbonate $\delta^{13}C$ record predicts release of roughly 10 to 20 times the present atmospheric oxygen inventory[52]. Recent findings suggest that oxygen was indeed very high during the Lomagundi excursion, including estimates of high sulphate and trace-metal levels in the ocean[53,55,56]. Equally tantalizing are suggestions of a precipitous drop in oxygen after the Lomagundi excursion[56,57]. The reasons for this rise and fall remain unresolved, although some models

blame extreme weathering of crust that developed under the generally O_2-lean Archaean atmosphere. This crust was rich in pyrite, which, when oxidized, would produce acidity and enhance delivery of key nutrients—phosphorus in particular[57]. Independent of the mechanism, this inferred nonlinear, reversible increase in atmospheric oxygen after the GOE stands in stark contrast to the classic models invoking unidirectional oxygen rise (Fig. 1). Few data are currently available, but no strong biotic response to these large-scale redox fluctuations has been recognized.

Oxygen and life during Earth's middle age

In the late 1990s, few grasped the full rise and fall of O_2 that may be captured in the Lomagundi excursion, but in a seminal paper published in 1998, Canfield[21] set the tone for the ensuing consequences by modelling a persistence of low marine oxygen conditions throughout the mid-Proterozoic from roughly 1.8 to 0.8 Gyr ago—long after the GOE. He went a step further and suggested pervasive euxinia in the deep ocean. (Euxinia refers to waters free of oxygen and rich in hydrogen sulphide, H_2S, like those that characterize the Black Sea today.) Whether he intended it or not, that view soon became one of a globally euxinic 'Canfield' ocean that dominated Earth's middle age. Some years later, many researchers, including Canfield, struggled to define a combination of factors, particularly the controls on primary production that would have sustained euxinia across such large expanses of the open ocean[58–60].

Nevertheless, building on the idea of ocean-scale euxinia, Anbar and Knoll[61] presented an intriguing thought experiment: because important micronutrients such as Mo are readily scavenged from sea water in the presence of hydrogen sulphide, might the mid-Proterozoic ocean have been broadly limited in these key metals, which are required enzymatically for the fixation and utilization of nitrogen? In today's oxic world, iron limits primary production in vast parts of the ocean, while Mo abounds. The situation may have been reversed under the low-oxygen conditions of the mid-Proterozoic. This nutrient state would have throttled the early diversity, distribution and abundances of eukaryotes—an idea explored later through phylogenomic analysis of protein structures and the implied histories of metal utilization in prokaryotic and eukaryotic organisms[62]. Scott *et al.*[59] found evidence for the hypothesized Mo deficiency in the mid-Proterozoic ocean (Box 2). Importantly, though, the observed Mo drawdown and complementary Mo isotope data[63] are inconsistent with anything close to ocean-wide euxinia.

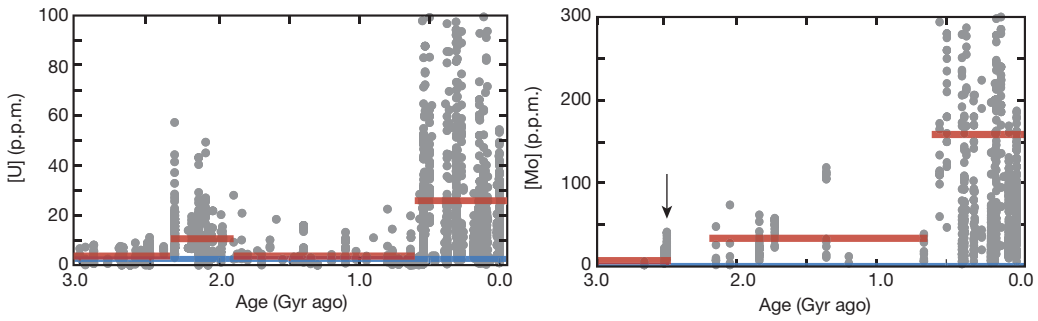
In the years following the initial excitement about mid-Proterozoic ocean-scale euxinia, a more nuanced and realistic model for ocean–atmosphere redox emerged. Oxygen was probably persistently or transiently very low in the atmosphere, perhaps even less than 0.1% of that present today (Fig. 1). For example, the apparent loss of manganese (Mn) from some mid-Proterozoic soils (palaeosols) opens up the possibility of markedly low atmospheric oxygen concentrations in the mid-Proterozoic well after the GOE[64]. Sedimentary chromium (Cr) isotope relationships[65] may, similarly, suggest limited terrestrial Mn oxidation for periods of the mid-Proterozoic hundreds of millions of years after the GOE. In modern environments, by analogy, Mn oxidation can proceed

BOX 2

Trace-element records of ocean redox evolution

Because the burial of redox-sensitive elements (RSEs) in marine sediments is greatly enhanced in anoxic settings, pervasive marine anoxia will result in RSE depletion in sea water. Further, the magnitude of enrichment of a given RSE in a local anoxic setting should scale with its marine reservoir size[104]. Large sedimentary RSE enrichments in local anoxic environments will only develop in a world with broadly oxic oceans (as on the modern Earth), whereas pervasively anoxic conditions will lead to decreased RSE reservoir sizes and thus muted sedimentary enrichments. If the redox state of the overlying water column can be independently constrained, then the magnitude of sedimentary RSE enrichments can be used to shed light on RSE reservoir size and thus global redox structure. Building from modern marine element mass balances and combining elements that respond to the presence of sulphidic conditions (Mo) with those that respond to anoxia with or without sulphide (U, Cr) it is possible to estimate the global redox landscape (percentage of various seafloor redox states, for example, anoxic, oxic, euxinic) using RSE data from locally anoxic environments (see, for instance, ref. 60).

Although this is a well-grounded approach, it is important to note that other factors (for example, organic fluxes, sulphide levels and bulk sediment accumulation rates) can affect the removal rate of a given RSE. These secondary effects translate into some degree of uncertainty in quantitative estimates; however, these should generally be minor relative to the robust first-order trends in RSE enrichment that we observe (Box 2 Figure) and the much greater errors associated with past practices of extrapolating redox conditions at single locations to the global ocean.



Box 2 Figure | Trace-metal records of evolving ocean redox conditions. Data have been filtered by independent methods to represent anoxic (left) and euxinic (anoxic and sulphidic; right) marine environments. Blue bars represent the range for upper continental crust. Red bars denote the average values for Archaean, early Proterozoic (left only), mid-Proterozoic and Neoproterozoic–Phanerozoic data. Data are from refs 56, 59, 60, 80. Large Phanerozoic U and Mo enrichments point to a dominantly oxic ocean (with low enrichments being linked predominantly to anoxic events or severe isolation). Large U enrichments in the early Proterozoic similarly suggest a well-oxygenated ocean, while persistently muted U enrichments in the mid-Proterozoic suggest the reversion back to a poorly ventilated ocean. Modest Mo enrichments in the mid-Proterozoic, however, suggest that only a moderate extent of this poorly oxygenated ocean was euxinic. The presence of significant Mo enrichments in the Archaean (arrow) suggests the presence of oxidative processes at least as far back as 2.5 Gyr ago[9].

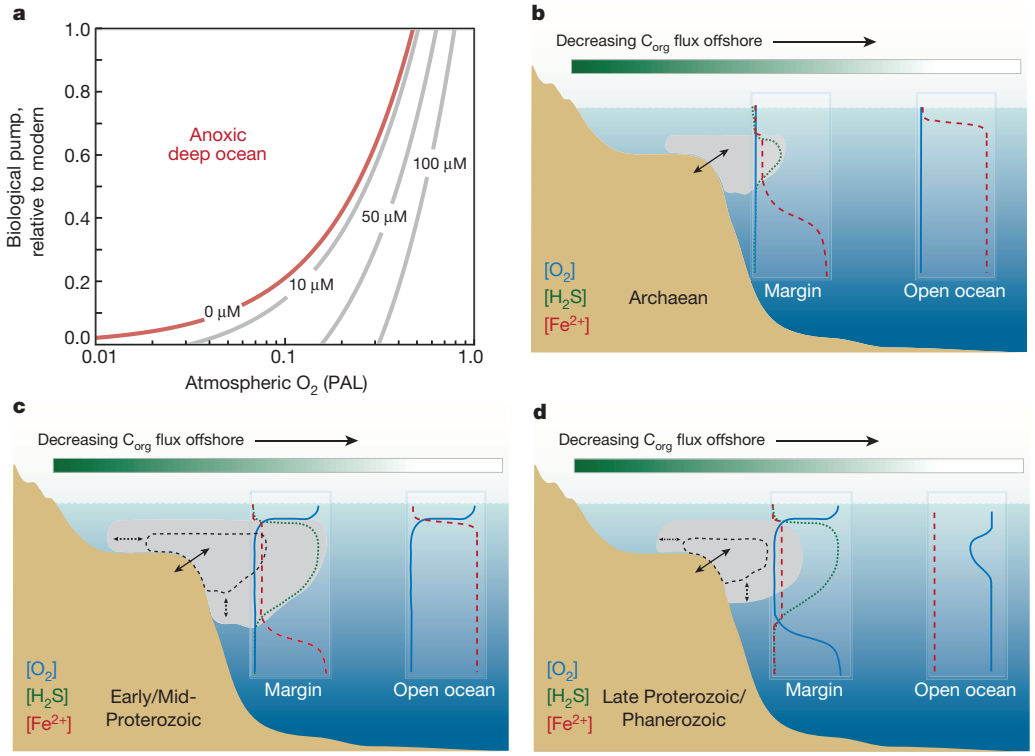


Figure 3 | Ocean ventilation and evolving ocean redox structure. **a**, Contours of globally averaged deep ocean O_2, which is largely set by a balance between O_2 introduced from the atmosphere and the respiration of settling organic matter in the ocean (the 'biological pump'). Calculations are performed as in Canfield[21] and Sarmiento *et al.*[98] but are recast in terms of atmospheric O_2 levels and carbon fluxes through the biological pump (both normalized to the modern Earth). Grey contours reflect globally averaged deep ocean O_2 concentration (in µM), with the red contour showing the boundary below which the modelled deep ocean becomes anoxic. **b–d**, Summary of an emerging model for the evolving first-order redox structure of the ocean (see text): **b**, Archaean; **c**, early/mid-Proterozoic; **d**, late Proterozoic/Phanerozoic. Left and right insets in each panel **b–d** show average profiles of O_2 (blue), H_2S (green) and Fe^{2+} (red); also shown (colour bar) is the general offshore decrease in local organic carbon (C_{org}) fluxes and its impact on the redox profile of the water column. Double-headed arrows denote expected expansion and contraction of sulphidic and/or ferruginous conditions (grey shading) along the productive and correspondingly reducing ocean margins. We emphasize that the Ediacaran, and much of the late Proterozoic more broadly, was most likely to have been marked by transient oscillation between states depicted in **c** and **d**. It is also important to note that small amounts of oxygen were probably present, locally and perhaps transiently, in the Archaean atmosphere and shallow ocean (**b**), perhaps as local oxygen oases for the latter[29].

rapidly at oxygen levels equivalent to $<10^{-3}$ PAL[66]—which would potentially place mid-Proterozoic atmospheric O_2 well below the commonly cited estimates based on traditional palaeosol work and assumptions of a persistently anoxic deep ocean (>1 to <40% PAL, respectively; Figs 1, 3a)[21,67]. Coupled ancient Cr-Mn cycling and our ability to extrapolate modern natural and experimental systems to quantify those ancient pathways precisely are active areas of research, as are the feedbacks necessary to modulate atmospheric O_2 at such low levels after its initial rise. Moreover, additional records of metal cycling on land through the Proterozoic will probably allow us to constrain better the timing and causes of increases in ocean and atmospheric oxygen contents that mark the shift to a very different late Proterozoic world.

Newer data emphasizing detailed iron speciation within shales suggested that the deep ocean remained dominantly anoxic[68], as Canfield[21] predicted, in response to the still low oxygen values in the atmosphere. But unlike the classic 'Canfield' euxinic ocean, the limited data are best explained by mostly iron-rich anoxic conditions with euxinia largely limited to biologically productive ocean margins and restricted marginal basins[59,69–72]. Today, organic productivity is highest in zones of nutrient upwelling along continental margins, and we can imagine the same situation in the early ocean—much like oxygen-minimum zones in the modern world (Fig. 3b–d). Decay of that settling organic matter removes oxygen from the deeper waters, and the generally low O_2 conditions of the mid-Proterozoic would have exacerbated those deficiencies (Fig. 3a). Persistent and pervasive low-oxygen conditions in the ocean and atmosphere might also have been favoured by copious anoxygenic photosynthesis linked to microbial iron and/or H_2S oxidation in the shallow ocean[73].

Recognizing the likelihood of a more redox stratified mid-Proterozoic ocean was a major step forward but unfortunately the 'proof' resided mostly with very broad extrapolations of inferred conditions at only a few locations. The risk is not unlike surmising the global redox state of the modern ocean through measurements along the highly productive upwelling region off Peru–Chile or within the nearly isolated anoxic Black Sea. The call was out for new approaches.

In response to concerns about over-extrapolation, combined elemental measurements and mass balance modelling is now permitting first-order spatial estimates for conditions across the full extent of sea floor, including those portions long-since lost to subduction, while also providing a more direct measure of the elemental abundances in sea water[60]. For example, Cr and Mo, because of their differing sensitivities to H_2S-free conditions, constrain ocean anoxia to at least 30–40% of the sea floor, and very possibly much more, for large intervals of the mid-Proterozoic, with the likelihood of elevated levels of dissolved iron (Box 2). Those portions of the deep ocean that were not fully anoxic may well have contained only trace levels of oxygen, a condition often referred to as 'suboxic'[69,74]. Euxinic waters, defined by the presence of H_2S, were potentially common enough to pull the concentrations of some key bioessential metals below those favoured by prokaryotes and eukaryotes[60,75], even if limited to only ~1–10% of the sea floor[60] (relative to ≪1% today). Specifically, there may have been persistent molybdenum-nitrogen co-limitation linked to euxinia through much of the mid-Proterozoic, and those molybdenum deficiencies ultimately may have played a major role in limiting the extent of euxinia[58]. Although considered to be less efficient, enzymatic pathways other than Mo-based

nitrogen fixation must also be considered in future studies. Furthermore, we cannot exclude the possibility of a very different phosphorus cycle at that time and lower-than-modern average phosphorus concentrations. Overall, a comprehensive network of nutrient-based feedbacks may have sustained oxygen at low levels with commensurate effects on marine life, including severe limits on eukaryote diversity and abundance. At the heart of these feedbacks were coupled rising and falling organic production, H_2S generation and metal availability within a relatively narrow range—as expressed in the famously 'boring' mid-Proterozoic $\delta^{13}C$ data, which are marked by exceptional consistency through time (Fig. 2).

Importantly, both modelled and measured evidence are lining up in favour of dominantly ferruginous, or iron-rich, conditions in the deep ocean through the Proterozoic[60,70,71], much like the earlier Archaean. An important implication is that the temporal distribution of economic-grade iron formations must reflect something other than just the redox state of the deep ocean—probably episodes of heightened plume activity within the mantle[76] and/or periods with higher iron concentrations in the hydrothermal fluids released on the sea floor[77]. Only near the end of the Proterozoic did oxygen take a big step up again, perhaps in response to first-order shifts in global-scale tectonics and glaciations in combination with biological innovations.

Another step towards the modern world

Despite a new wave of excellent work, much remains unknown about the redox structure of the ocean and atmosphere during the later part of the Proterozoic (formally known as the Neoproterozoic) between roughly 0.8 and 0.55 Gyr ago and its relationship with evolving life. This gap is a bit surprising given its relatively young age, the comparatively good quality and quantity of available rocks to study, and the abundant recent work on this interval. Yet, the common interpretations tread close to a worrisome circularity: the emergence of animals is typically attributed to a second big O_2 step long after the GOE (a so-called Neoproterozoic Oxidation Event[78]), but animals are just as often cited as evidence for the oxygenation. Other signs of Neoproterozoic oxygenation lie with evidence for deep marine O_2 (refs 79, 80) and problematic explanations for Earth's greatest negative carbon isotope excursion (Fig. 2)—the so-called Shuram-Wonaka anomaly[81,82], which is interpreted to be of either primary or secondary origin[83] (reviewed in ref. 82). Other data point instead, in seeming contradiction, to a persistence of expansive anoxic (iron-rich, that is, ferruginous, and euxinic) marine waters[84].

Amidst the apparent confusion, new research is steering us towards consistent threads that run through all these data by invoking anoxic conditions on productive late Neoproterozoic ocean margins and oxygenation, at least episodically, in the deeper waters (Fig. 3c, d). Indeed, some of the available trace-metal data point to very low extents of euxinic and ferruginous waters at times during the latest Neoproterozoic—also known as the Ediacaran (~635–542 Myr ago)—potentially in phase with major shifts in eukaryotic/animal innovation (reviewed in ref. 85; Box 2). However, we also expect large-scale temporal variability in marine redox conditions, and climate/glaciation may have been a driver of biogeochemical destabilization and a key factor behind the escape from the oxygen-lean stasis that characterized the mid-Proterozoic[86,87]. For instance, one can imagine that shifts in nutrient cycles at the end of the Marinoan 'Snowball Earth' glaciation, the second of two major ice ages in the Neoproterozoic, may have triggered the organic productivity/burial that then spawned the rise in oxygen in the early Ediacaran[80], and trace-metal enrichments suggest a widely oxygenated ocean at about 630 and 550 Myr ago[59,80]. The detailed timing and persistence of O_2 accumulation in the Neoproterozoic ocean and the transition into the younger Phanerozoic are not well known and allow for rising and falling oxygen concentrations during the Ediacaran, as well as the possibility of earlier, even pre-Snowball Earth, oxygenation that may have helped trigger the climate events that followed. It is also likely that shifts in global tectonics during the Neoproterozoic played a strong role in initiating late-Proterozoic global environmental change. Continuous diversification of algae (eukaryotic primary producers) throughout the Neoproterozoic may also have helped to initiate late-Proterozoic global environmental change by altering basic aspects of the marine carbon cycle.

Little is known about the specific relationship between early animals and oxygen. The earliest animals were sponges or sponge-grade[88–90], and their small sizes and high rates of internal ventilation suggest that they may have had relatively low oxygen demand. If one is inclined to link the rise of animals to a rise of oxygen, a logical corollary is that atmospheric oxygen during the preceding mid-Proterozoic must have been at least transiently very low to explain the apparent lack of animals—maybe (much) less than 1% of today's level (Fig. 1). Butterfield[91] suggested instead that the generally concurrent rise of animals and oxygen was mostly a coincidence or, alternatively, that animal evolution itself triggered the oxygenation event. By this argument, the long delay in animal emergence reflects instead the intrinsic timescales of evolution and the complexity of gene expression and cell signalling in animals, consistent with the apparent lack of animals during the much earlier O_2-rich Lomagundi excursion. Others researchers assert various scenarios that demand oxygen in appreciable amounts[88] to explain high animal diversity, large mobile bilaterians, the advent of biomineralization (skeletons), wide niche expansion including habitats below the sea floor, and complex predator–prey relationships[92]. At the same time, we know that animals will alter ecosystem structure and profoundly influence the carbon cycle[88,93], and thus local and broader oxygen levels, by burrowing into sediments, for example. In every case, environment and co-evolving life participate in myriad feedback loops, wherein changes to one generally affect the other. Thus, we warn against end-member arguments in this debate.

The way forward

Informed by increasing sophistication in elemental and isotopic proxy approaches, we can now say with much greater confidence when and why the redox structure of the ocean and atmosphere varied through time. Through this window, we can view an ocean and atmosphere that were mostly oxygen-starved for almost 90% of Earth's history.

So what are the next great opportunities in studies of early oxygen? Of particular value are proxies for seawater composition and linked numerical models that make it possible to extrapolate beyond local conditions and allow, perhaps for the first time, access to the chemical landscape of the ocean as a whole. We recall that the goal is to characterize conditions on a sea floor that is mostly lost through subduction, and the records that we do have from the ancient ocean margin are intrinsically vulnerable to local controls, such as basin restriction and elevated local levels of primary production. We also need additional quantitative tracers of oxygen levels in the atmosphere, given how hard it is to quantify its composition with confidence using mostly oxygen levels inferred for the ocean. And despite significant steps forward, too little is known about the precise timing of the emergence of oxygenic photosynthesis. In this search, organic and inorganic geochemical methods must be used with full awareness of all possibilities of overprinting and contamination. As always, novel approaches applied to more and better samples with the strongest possible age and sedimentological controls will continue to drive the research, with the latter providing independent constraints on depositional conditions that complement geochemical analysis.

The Proterozoic is book-ended by the two greatest geobiological events in Earth's history—the GOE and the dramatic changes among life and environment in the late Neoproterozoic—and these will continue to grab much of the attention. Armed with a better grasp of the history of oxygenic photosynthesis and the full range of evolving oxygen-consuming reactions as tied to processes both on and deep within the Earth, we will correctly tackle the first rise of atmospheric oxygen as the complicated, protracted, dynamic process that it must have been. Refined views of the history of continent formation will inform these discussions.

The billion or more years of history beyond the initial oxygenation of the atmosphere will remain a prime target, particularly given recent suggestions of a remarkable persistence of mostly very low oxygen levels, perhaps more akin to the Archaean than the modern world, and their strangle-hold on early complex life. Full resolution of the feedbacks involved will be a great

leap forward. Finally, researchers will ask more and better questions about the unique confluence of global-scale climatic, evolutionary and tectonic events that once and for all broke the cycle of low oxygen on Earth, less than a billion years ago, and set the stage for everything that followed, including the emergence of animal life. Increasingly within that mix may be indications of dramatic Neoproterozoic oxygenation well before the Ediacaran[94], and even the 'Snowball Earth' glaciations, thus challenging us to unravel the complex cause-and-effect relationships. And we should not forget that just as environmental change can drive the evolution of life, the reverse is also true. A few billion years after the earliest life, the evolutionary clock may also have been timed just right for big change.

Finally, we summarize the changing understanding of the GOE. In 2002, Holland[2] coined the term 'Great Oxidation Event' to formalize the concept that had emerged long before—that the atmosphere shifted from being fundamentally reducing to oxidizing over an interval from roughly 2.4 to 2.1 Gyr ago. The presumed disappearance of NMD sulphur isotope signals narrowed that window to between 2.4 and 2.3 Gyr ago[35]. No doubt a fundamental shift did occur over this general interval as part of a much broader, long-term progression towards higher amounts of oxygen. But equally certain now is that biospheric oxygen did not follow the simple unidirectional, step-punctuated rise traditionally envisioned[95]. Instead, imagine something more like a roller coaster ride, with dynamic rising and falling oxygen levels in the ocean and atmosphere—starting perhaps as early as 3.0 Gyr ago—superimposed on a first-order trend from generally low to intermediate to high concentrations over a period of perhaps two and half billion years. In this light, the Great Oxidation Event was a transition (a Great Oxygen Transition or GOT, perhaps), more protracted and dynamic than event-like. And any assertions of greatness, particularly those tied specifically to the apparent loss of NMD sulphur isotope signals, may undersell the importance of oxygen variability that came well before and long after an isotopic milestone perhaps blurred by sedimentary recycling and complicated by processes not yet discovered. But 'great' works if we think longer-term, fundamental redox shift, and no matter how we define the GOE, the 'how, when and why' behind Earth's dynamic oxygen history will continue to motivate a generation of researchers.

Received 11 April 2013; accepted 21 January 2014.

1. Roscoe, S. M. Huronian rocks and uraniferous conglomerates in the Canadian Shield. *Geol. Surv. Pap. Can.* 68–40 (1969).
2. Holland, H. D. Volcanic gases, black smokers, and the Great Oxidation Event. *Geochim. Cosmochim. Acta* **66,** 3811–3826 (2002).
Formalized the notion of the GOE and highlighted the important balance between oxygen production and oxygen-buffering reactions, via reduced volatile compounds, in modulating the prevailing redox state at Earth's surface.
3. Holland, H. D. The oxygenation of the atmosphere and oceans. *Phil. Trans. R. Soc. B* **361,** 903–915 (2006).
4. Canfield, D. E. The early history of atmospheric oxygen: Homage to Robert M. Garrels. *Annu. Rev. Earth Planet. Sci.* **33,** 1–36 (2005).
5. Ohmoto, H., Watanabe, Y., Ikemi, H., Poulson, S. R. & Taylor, B. E. Sulfur isotope evidence for an oxic Archaean atmosphere. *Nature* **442,** 908–911 (2006).
6. Brocks, J. J., Logan, G. A., Buick, R. & Summons, R. E. Archean molecular fossils and the early rise of eukaryotes. *Science* **285,** 1033–1036 (1999).
Essential organic biomarker study that provided the most-cited evidence for the earliest records of oxygen-producing photosynthesis, well before the GOE; the integrity of the biomarker data has been challenged in recent years.
7. Brocks, J. J. Millimeter-scale concentration gradients of hydrocarbons in Archean shales: Live-oil escape or fingerprint of contamination? *Geochim. Cosmochim. Acta* **75,** 3196–3213 (2011).
8. Rasmussen, B., Fletcher, I. R., Brocks, J. J. & Kilburn, M. R. Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature* **455,** 1101–1104 (2008).
9. Anbar, A. D. *et al.* A whiff of oxygen before the Great Oxidation Event? *Science* **317,** 1903–1906 (2007).
Drew attention to the possibility of oxidative weathering of the continents—well before the GOE; recent challenges to the late Archaean organic biomarker record have elevated the value of the study's inorganic data as likely signatures of pre-GOE oxygenesis.
10. Gaillard, F., Scaillet, B. & Arndt, N. T. Atmospheric oxygenation caused by a change in volcanic degassing pressure. *Nature* **478,** 229–232 (2011).
11. Kump, L. R. & Barley, M. E. Increased subaerial volcanism and the rise of atmospheric oxygen 2.5 billion years ago. *Nature* **448,** 1033–1036 (2007).
12. Catling, D. C., Zahnle, K. J. & McKay, C. P. Biogenic methane, hydrogen escape, and the irreversible oxidation of early Earth. *Science* **293,** 839–843 (2001).
Model exploring the consequences of atmospheric hydrogen escape for the redox budget of the evolving Earth; it has become a crucial lynchpin in the examination of Earth's oxygenation within a planetary context.
13. Claire, M. W., Catling, D. C. & Zahnle, K. J. Biogeochemical modelling of the rise in atmospheric oxygen. *Geobiology* **4,** 239–269 (2006).
14. Zahnle, K. J., Caltling, D. C. & Claire, M. W. The rise of oxygen and the hydrogen hourglass. *Chem. Geol.* (in the press).
15. Kirschvink, J. L. & Kopp, R. E. Paleoproterozic icehouses and the evolution of oxygen mediating enzymes: the case for a late origin of Photosystem-II. *Phil. Trans. R. Soc. B* **363,** 2755–2765 (2008).
16. Farquhar, J., Bao, H. & Thiemens, M. Atmospheric influence of Earth's earliest sulfur cycle. *Science* **289,** 756–758 (2000).
Arguably the 'smoking gun' for the GOE—the loss of non-mass-dependent sulphur isotope fractionations—and thus launched a new wave of sulphur studies in Precambrian biogeochemistry and refined our understanding of early oxygenation.
17. Pavlov, A. A. & Kasting, J. F. Mass-independent fractionation of sulfur isotopes in Archean sediments: strong evidence for an anoxic Archean atmosphere. *Astrobiology* **2,** 27–41 (2002).
18. Zahnle, K. J., Claire, M. & Catling, D. The loss of mass-independent fractionation in sulfur due to a Paleoproterozoic collapse of atmospheric methane. *Geobiology* **4,** 271–283 (2006).
19. Zerkle, A. L., Claire, M. W., Domagal-Goldman, S. D., Farquhar, J. & Poulton, S. W. A bistable organic-rich atmosphere on the Neoarchaean Earth. *Nature Geosci.* **5,** 359–363 (2012).
20. Halevy, I., Johnston, D. T. & Schrag, D. P. Explaining the structure of the Archean mass-independent sulfur isotope record. *Science* **329,** 204–207 (2010).
21. Canfield, D. E. A new model for Proterozoic ocean chemistry. *Nature* **396,** 450–453 (1998).
Spawned the concept of the 'Canfield' ocean by developing the idea that the ocean remained anoxic and probably euxinic for a billion years of the mid-Proterozoic, thus highlighting the essential lag between atmospheric and oceanic oxygenation and setting the stage for a generation of research in Precambrian oxygenation.
22. Rosing, M. T. & Frei, R. U-rich Archaean sea-floor sediments from Greenland—indications of 3700 Ma oxygenic photosynthesis. *Earth Planet. Sci. Lett.* **217,** 237–244 (2004).
23. Cairns-Smith, A. G. Precambrian solution photochemistry, inverse segregation, and banded iron formations. *Nature* **276,** 807–808 (1978).
24. Crowe, S. A. *et al.* Photoferrotrophs thrive in an Archean ocean analogue. *Proc. Natl Acad. Sci. USA* **105,** 15938–15943 (2008).
25. Konhauser, K. O. *et al.* Could bacteria have formed the Precambrian banded iron formations? *Geology* **30,** 1079–1082 (2002).
26. Bosak, T., Knoll, A. H. & Petroff, A. P. The meaning of stromatolites. *Annu. Rev. Earth Planet. Sci.* **41,** 21–44 (2013).
27. Waldbauer, J. R., Newman, D. K. & Summons, R. E. Microaerobic steroid biosynthesis and the molecular fossil record of Archean life. *Proc. Natl Acad. Sci. USA* **108,** 13409–13414 (2011).
28. French, K. L. *et al.* Archean hydrocarbon biomarkers: Archean or not? *Goldschmidt 2013 Conf. Abstr.* http://goldschmidtabstracts.info/2013/1110.pdf (2013).
29. Kasting, J. F. in *The Proterozoic Biosphere* (eds Schopf, J. W. & Klein, C.) Ch. 26.2 1185–1188 (Cambridge Univ. Press, 1992).
30. Farquhar, J., Zerkle, A. L. & Bekker, A. Geological constraints on the origin of oxygenic photosynthesis. *Photosynth. Res.* **107,** 11–36 (2011).
31. Kendall, B., Creaser, R. A., Gordon, G. W. & Anbar, A. D. Re-Os and Mo isotope systematics of black shales from the Middle Proterozoic Velkerri and Wollogorang Formations, McArthur Basin, northern Australia. *Geochim. Cosmochim. Acta* **73,** 2534–2558 (2009).
32. Reinhard, C. T., Raiswell, R., Scott, C., Anbar, A. D. & Lyons, T. W. A late Archean sulfidic sea stimulated by early oxidative weathering of the continents. *Science* **326,** 713–716 (2009).
33. Reinhard, C. T., Planavsky, N. J. & Lyons, T. W. Long-term sedimentary recycling of rare sulphur isotope anomalies. *Nature* **497,** 100–103 (2013).
34. Guo, Q. *et al.* Reconstructing Earth's surface oxidation across the Archean-Proterozoic transition. *Geology* **37,** 399–402 (2009).
35. Bekker, A. *et al.* Dating the rise of atmospheric oxygen. *Nature* **427,** 117–120 (2004).
First study to attempt to fingerprint the GOE precisely, using a tightly constrained stratigraphic record of the disappearance of NMD sulphur isotope fractionations, thus defining a temporal context for oxygenation models and major related climate events.
36. Konhauser, K. O. *et al.* Aerobic bacterial pyrite oxidation and acid rock drainage during the Great Oxidation Event. *Nature* **478,** 369–373 (2011).
37. Crowe, S. *et al.* Atmospheric oxygenation three billion years ago. *Nature* **501,** 535–538 (2013).
38. Berner, R. A. *The Phanerozoic Carbon Cycle* (Oxford Univ. Press, 2004).
39. Goldblatt, C., Lenton, T. M. & Watson, A. J. Bistability of atmospheric oxygen and the Great Oxidation. *Nature* **443,** 683–686 (2006).
40. Canil, D. Vanadium in peridotites, mantle redox and tectonic environments: Archean to present. *Earth Planet. Sci. Lett.* **195,** 75–90 (2002).
41. Li, Z. X. A. & Lee, C. T. A. The constancy of upper mantle fO_2 through time inferred from V/Sc ratios in basalts. *Earth Planet. Sci. Lett.* **228,** 483–493 (2004).
42. Trail, D., Watson, E. B. & Tailby, N. D. The oxidation state of Hadean magmas and implications for early Earth's atmosphere. *Nature* **480,** 79–82 (2011).
43. Konhauser, K. O. *et al.* Oceanic nickel depletion and a methanogen famine before the Great Oxidation Event. *Nature* **458,** 750–753 (2009).

44. Evans, D. A., Beukes, N. J. & Kirschvink, J. L. Low-latitude glaciation in the Palaeoproterozoic era. *Nature* **386,** 262–266 (1997).
45. Habicht, K. S., Gade, M., Thamdrup, B., Berg, P. & Canfield, D. E. Calibration of sulfate levels in the Archean ocean. *Science* **298,** 2372–2374 (2002).
46. Haqq-Misra, J. D., Domagal-Goldmann, S. D., Kasting, P. J. & Kasting, J. F. A revised, hazy methane greenhouse for the Archean Earth. *Astrobiology* **8,** 1127–1137 (2008).
47. Jamieson, J. W., Wing, B. A., Farquhar, J. & Hannington, M. D. Neoarchaean seawater sulphate concentrations from sulphur isotopes in massive sulphide ore. *Nature Geosci.* **6,** 61–64 (2013).
48. Pavlov, A. A., Kasting, J. F. & Brown, L. L. Greenhouse warming by CH_4 in the atmosphere of early Earth. *J. Geophys. Res.* **105,** 11981–11990 (2000).
49. Knittel, K. & Boetius, A. Anaerobic oxidation of methane: progress with an unknown process. *Annu. Rev. Microbiol.* **63,** 311–334 (2009).
50. Gough, D. O. Solar interior structure and luminosity variations. *Sol. Phys.* **74,** 21–34 (1981).
51. Sagan, C. & Mullen, G. Earth and Mars: evolution of atmospheres and surface temperatures. *Science* **177,** 52–56 (1972).
52. Karhu, J. A. & Holland, H. D. Carbon isotopes and the rise of atmospheric oxygen. *Geology* **24,** 867–870 (1996).
53. Planavsky, N. J., Bekker, A., Hofmann, A., Owens, J. D. & Lyons, T. W. Sulfur record of rising and falling marine oxygen and sulfate levels during the Lomagundi event. *Proc. Natl Acad. Sci. USA* **109,** 18300–18305 (2012).
54. Hayes, J. M. & Waldbauer, J. R. The carbon cycle and associated redox processes through time. *Phil. Trans. R. Soc. B* **361,** 931–950 (2006).
55. Schröder, S., Bekker, A., Beukes, N. J., Strauss, H. & van Niekerk, H. S. Rise in seawater sulphate concentration associated with the Paleoproterozoic positive carbon isotope excursion: evidence from sulphate evaporites in the ~2.2–2.1 Gyr shallow-marine Lucknow Formation, South Africa. *Terra Nova* **20,** 108–117 (2008).
56. Partin, C. A. *et al.* Large-scale fluctuations in Precambrian atmospheric and oceanic oxygen levels from the record of U in shales. *Earth Planet. Sci. Lett.* **369–370,** 284–293 (2013).
57. Bekker, A. & Holland, H. D. Oxygen overshoot and recovery during the early Paleoproterozoic. *Earth Planet. Sci. Lett.* **317–318,** 295–304 (2012).
58. Boyle, R. A. *et al.* Nitrogen cycle feedbacks as a control on euxinia in the mid-Proterozoic ocean. *Nature Commun.* **4,** 1533 (2013).
59. Scott, C. *et al.* Tracing the stepwise oxygenation of the Proterozoic biosphere. *Nature* **452,** 456–459 (2008).
60. Reinhard, C. *et al.* Proterozoic ocean redox and biogeochemical stasis. *Proc. Natl Acad. Sci. USA* **110,** 5357–5362 (2013).
State-of-the-art exploration of the redox landscape of the mid-Proterozoic ocean—with important implications for the mechanisms behind the persistently low levels of biospheric oxygen that defined the 'boring billion'.
61. Anbar, A. D. & Knoll, A. H. Proterozoic ocean chemistry and evolution: a bioinorganic bridge? *Science* **297,** 1137–1142 (2002).
Building from the concept of the 'Canfield' ocean, this was the first paper to develop the idea of possible trace-metal limitations under assumed widespread euxinia in the mid-Proterozoic ocean as a throttle on early eukaryotic expansion.
62. Dupont, C. L., Butcher, A., Valas, R. E., Bourne, P. E. & Caetano-Anolles, G. History of biological metal utilization inferred through phylogenomic analysis of protein structures. *Proc. Natl Acad. Sci. USA* **107,** 10567–10572 (2010).
63. Arnold, G. L., Anbar, A. D., Barling, J. & Lyons, T. W. Molybdenum isotope evidence for widespread anoxia in Mid-Proterozoic oceans. *Science* **304,** 87–90 (2004).
64. Zbinden, E. A., Holland, H. D., Feakes, C. R. & Dobos, S. K. The Sturgeon Falls paleosol and the composition of the atmosphere 1.1 Ga Bp. *Precambr. Res.* **42,** 141–163 (1988).
65. Frei, R., Gaucher, C., Poulton, S. W. & Canfield, D. E. Fluctuations in Precambrian atmospheric oxygenation recorded by chromium isotopes. *Nature* **461,** 250–253 (2009).
66. Clement, B. G., Luther, G. W. & Tebo, B. M. Rapid, oxygen-dependent microbial Mn(II) oxidation kinetics at sub-micromolar oxygen concentrations in the Black Sea suboxic zone. *Geochim. Cosmochim. Acta* **73,** 1878–1889 (2009).
67. Rye, R. & Holland, H. D. Paleosols and the evolution of atmospheric oxygen: a critical review. *Am. J. Sci.* **298,** 621–672 (1998).
68. Shen, Y., Canfield, D. E. & Knoll, A. H. Middle Proterozoic ocean chemistry: Evidence from McArthur Basin, Northern Australia. *Am. J. Sci.* **302,** 81–109 (2002).
69. Lyons, T. W., Anbar, A. D., Severmann, S., Scott, C. & Gill, B. C. Tracking euxinia in the ancient ocean: A multiproxy perspective and Proterozoic case study. *Annu. Rev. Earth Planet. Sci.* **37,** 507–534 (2009).
70. Planavsky, N. J. *et al.* Widespread iron-rich conditions in the mid-Proterozoic ocean. *Nature* **477,** 448–451 (2011).
71. Poulton, S. W. & Canfield, D. E. Ferruginous conditions: a dominant feature of the ocean through Earth's history. *Elements* **7,** 107–112 (2011).
72. Poulton, S. W., Fralick, P. W. & Canfield, D. E. Spatial variability in oceanic redox structure 1.8 billion years ago. *Nature Geosci.* **3,** 486–490 (2010).
73. Johnston, D. T., Wolfe-Simon, F., Pearson, A. & Knoll, A. H. Anoxygenic photosynthesis modulated Proterozoic oxygen and sustained Earth's middle age. *Proc. Natl Acad. Sci. USA* **106,** 16925–16929 (2009).
74. Slack, J. F., Grenne, T., Bekker, A., Rouxel, O. J. & Lindberg, P. A. Suboxic deep seawater in the late Paleoproterozoic: evidence from hematitic chert and iron formation related to seafloor-hydrothermal sulfide deposits, central Arizona, USA. *Earth Planet. Sci. Lett.* **255,** 243–256 (2007).
75. Glass, J. B., Wolfe-Simon, F. & Anbar, A. D. Coevolution of metal availability and nitrogen assimilation in cyanobacteria and algae. *Geobiology* **7,** 100–123 (2009).
76. Bekker, A. *et al.* Iron formation: the sedimentary product of a complex interplay among mantle, tectonic, oceanic, and biospheric processes. *Econ. Geol.* **105,** 467–508 (2010).
77. Kump, L. R. & Seyfried, W. E. Hydrothermal Fe fluxes during the Precambrian: effect of low oceanic sulfate concentrations and low hydrostatic pressure on the composition of black smokers. *Earth Planet. Sci. Lett.* **235,** 654–662 (2005).
78. Och, L. M. & Shields-Zhou, G. A. The Neoproterozoic oxygenation event: Environmental perturbations and biogeochemical cycling. *Earth Sci. Rev.* **110,** 26–57 (2012).
79. Canfield, D. E., Poulton, S. W. & Narbonne, G. M. Late-Neoproterozoic deep-ocean oxygenation and the rise of animal life. *Science* **315,** 92–95 (2007).
80. Sahoo, S. K. *et al.* Ocean oxygenation in the wake of the Marinoan glaciation. *Nature* **489,** 546–549 (2012).
81. Fike, D. A., Grotzinger, J. P., Pratt, L. M. & Summons, R. E. Oxidation of the Ediacaran ocean. *Nature* **444,** 744–747 (2006).
82. Grotzinger, J. P., Fike, D. A. & Fischer, W. W. Enigmatic origin of the largest-known carbon isotope excursion in Earth's history. *Nature Geosci.* **4,** 285–292 (2011).
83. Swart, P. K. & Kennedy, M. J. Does the global stratigraphic reproducibility of $\delta^{13}C$ in Neoproterozoic carbonates require a marine origin? A Pliocene-Pleistocene comparison. *Geology* **40,** 87–90 (2012).
84. Canfield, D. E. *et al.* Ferruginous conditions dominated later Neoproterozoic deep-water chemistry. *Science* **321,** 949–952 (2008).
85. Lyons, T. W., Reinhard, C. T., Love, G. D. & Xiao, S. in *Fundamentals of Geobiology* (eds Knoll, A. H., Canfield, D. E. & Konhauser, K. O.) 371–402 (Blackwell, 2012).
86. Planavsky, N. *et al.* The evolution of the marine phosphate reservoir. *Nature* **467,** 1088–1090 (2010).
87. Swanson-Hysell, N. L. *et al.* Cryogenian glaciation and the onset of carbon-isotope decoupling. *Science* **328,** 608–611 (2010).
88. Erwin, D. H. *et al.* The Cambrian conundrum: early divergence and later ecological success in the early history of animals. *Science* **334,** 1091–1097 (2011).
Essential overview of our present understanding of the cause-and-effect relationships among early animal evolution and diversification, increasing ecological complexity, and environmental change—particularly oxygenation of the ocean and atmosphere.
89. Love, G. D. *et al.* Fossil steroids record the appearance of Demospongiae during the Cryogenian period. *Nature* **457,** 718–721 (2009).
90. Maloof, A. C. *et al.* Possible animal-body fossils in pre-Marinoan limestones from South Australia. *Nature Geosci.* **3,** 653–659 (2010).
91. Butterfield, N. J. Oxygen, animals and oceanic ventilation: an alternative view. *Geobiology* **7,** 1–7 (2009).
92. Sperling, E. A. Oxygen, ecology, and the Cambrian radiation of animals. *Proc. Natl Acad. Sci. USA* **110,** 13446–13451 (2013).
93. Logan, G. A., Hayes, J. M., Hieshima, G. B. & Summons, R. E. Terminal Proterozoic reorganization of biogeochemical cycles. *Nature* **376,** 53–56 (1995).
94. Baldwin, G. J., Nägler, T. F., Gerber, N. D., Turner, E. C. & Kamber, B. S. Mo isotopic composition of the mid-Neoproterozoic ocean: an iron formation perspective. *Precambr. Res.* **230,** 168–178 (2013).
95. Kump, L. R. The rise of atmospheric oxygen. *Nature* **451,** 277–278 (2008).
96. Berner, R. A. & Canfield, D. E. A new model for atmospheric oxygen over Phanerozoic time. *Am. J. Sci.* **289,** 333–361 (1989).
97. Bergman, N. M., Lenton, T. M. & Watson, A. J. COPSE: A new model of biogeochemical cycling over Phanerozoic time. *Am. J. Sci.* **304,** 397–437 (2004).
98. Sarmiento, J. L., Herbert, T. D. & Toggweiler, J. R. Causes of anoxia in the world ocean. *Glob. Biogeochem. Cycles* **2,** 115–128 (1988).
99. Overmann, J., Beatty, J. T., Krouse, H. R. & Hall, K. J. The sulfur cycle in the chemocline of a meromictic salt lake. *Limnol. Oceanogr.* **41,** 147–156 (1996).
100. Tromp, T. K., VanCappellen, P. & Key, R. M. A global model for the early diagenesis of organic carbon and organic phosphorus in marine sediments. *Geochim. Cosmochim. Acta* **59,** 1259–1284 (1995).
101. Kharecha, P., Kasting, J. & Siefert, J. A. coupled atmosphere-ecosystem model of the early Archean Earth. *Geobiology* **3,** 53–76 (2005).
102. Thunell, R. C. *et al.* Organic carbon fluxes, degradation, and accumulation in an anoxic basin: Sediment trap results from the Cariaco Basin. *Limnol. Oceanogr.* **45,** 300–308 (2000).
103. Messie, M. *et al.* Potential new production estimates in four eastern boundary upwelling ecosystems. *Prog. Oceanogr.* **83,** 151–158 (2009).
104. Algeo, T. J. & Lyons, T. W. Mo–total organic carbon covariation in modern anoxic marine environments: Implications for analysis of paleoredox and paleohydrographic conditions. *Paleoceanography* **21,** PA1016 (2006).

Acknowledgements Funding from NSF-EAR, the NASA Exobiology Program, the NASA Astrobiology Institute, and the Agouron Institute supported this work. C.T.R. acknowledges support from an O. K. Earl Postdoctoral Fellowship in Geological and Planetary Sciences at the California Institute of Technology. N.J.P. acknowledges support from NSF-EAR-PDF. Comments and criticism from A. Bekker, D. Erwin, I. Halevy and D. Johnston improved the manuscript. A. Bekker was helpful in discussions about the GOE and suggested the acronym 'GOT'.

Author Contributions C.T.R. and N.J.P. designed the model for O_2-producing photosynthesis and its relationship to Archaean organic carbon presented in Box 1. C.T.R. and N.J.P. compiled the database, and C.T.R. performed the modelling presented in Box 1. T.W.L. wrote the manuscript with major contributions from C.T.R. and N.J.P.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence should be addressed to T.W.L. (timothy.lyons@ucr.edu).

ARTICLE

doi:10.1038/nature13013

Geriatric muscle stem cells switch reversible quiescence into senescence

Pedro Sousa-Victor[1]†, Susana Gutarra[1]*, Laura García-Prat[1]*, Javier Rodriguez-Ubreva[2], Laura Ortet[1], Vanessa Ruiz-Bonilla[1], Mercè Jardí[1], Esteban Ballestar[2], Susana González[3], Antonio L. Serrano[1], Eusebio Perdiguero[1] & Pura Muñoz-Cánoves[1,4]

Regeneration of skeletal muscle depends on a population of adult stem cells (satellite cells) that remain quiescent throughout life. Satellite cell regenerative functions decline with ageing. Here we report that geriatric satellite cells are incapable of maintaining their normal quiescent state in muscle homeostatic conditions, and that this irreversibly affects their intrinsic regenerative and self-renewal capacities. In geriatric mice, resting satellite cells lose reversible quiescence by switching to an irreversible pre-senescence state, caused by derepression of p16^{INK4a} (also called Cdkn2a). On injury, these cells fail to activate and expand, undergoing accelerated entry into a full senescence state (geroconversion), even in a youthful environment. *p16^{INK4a}* silencing in geriatric satellite cells restores quiescence and muscle regenerative functions. Our results demonstrate that maintenance of quiescence in adult life depends on the active repression of senescence pathways. As p16^{INK4a} is dysregulated in human geriatric satellite cells, these findings provide the basis for stem-cell rejuvenation in sarcopenic muscles.

A hallmark of advanced ageing in humans is sarcopenia, an age-related loss of skeletal muscle mass and function, which is maximal in geriatric individuals and in patients with progeria syndromes[1]. Failure of the regeneration machinery of sarcopenic muscle to replace damaged myofibres is one of the major causes of the physical incapacitation and loss of independence in the geriatric population[2–4].

Skeletal muscle homeostasis and regeneration require a population of muscle-specific Pax7-expresing adult stem cells, named satellite cells, which are normally quiescent[5]. When stimulated by damage or stress, these G_0-arrested satellite cells are activated and enter the cell cycle, to proliferate further and form new fibres or self-renew to reconstitute the satellite cell pool[6,7]. Hetero-parabiosis and hetero-grafting experiments between young and (non-geriatric) old rodents have led to the prevalent view that the satellite-cell-dependent muscle regeneration decline with ageing can be reversed by exposure to an external youthful environment (reviewed in refs 3, 8–11). Notably, in very old (geriatric) mice with sarcopenia, we observed a marked drop in the muscle regenerative capacity compared to old non-sarcopenic mice. Whether this regenerative decline is ascribed to satellite cell intrinsic changes or to changes in the systemic environment is unknown.

Here, by analysing comparatively the gene expression program of quiescent satellite cells from mice at distinct ages, we demonstrate that geriatric satellite cells in homeostatic conditions undergo irreversible intrinsic changes that inhibit maintenance of the bona fide quiescence state, by switching to a senescence-like state. We have investigated the nature of these intrinsic alterations and their consequences on muscle regeneration at geriatric age in mice and humans, and their occurrence in progeria syndromes. Our findings indicate that, normally, the satellite cell protects its quiescence by mechanisms that inhibit irreversible cell-cycle withdrawal, thus avoiding senescence. This protective mechanism in stem cells is lost at advanced age.

Sharp regenerative decline of geriatric muscle stem cells

To delimit the period of maximal age-dependent muscle decline *in vivo* we analysed muscle properties in wild-type mice of 2–3 (young), 5–6 (adult), 20–24 (old) and 28–32 (geriatric) months of age. Compared to young/adult mice, fibre atrophy was detected at 20–24 months; however, loss of myofibre innervation, re-expression of embryonic myosin heavy chain (eMHC) and central nucleation were prominent at 28 months and thereafter, correlating with poor muscle performance (Extended Data Fig. 1a–d), indicating that sarcopenia occurs principally at geriatric age. Sarcopenia features were also evident in 12-month-old SAMP8 mice (a mouse model of progeria) compared to SAMR1 control mice[12] (Extended Data Fig. 1e).

We next analysed the regenerative capacity of satellite cells from physiologically ageing mice, with or without sarcopenia, after muscle injury by cardiotoxin (CTX) injection. Compared to adult mice, regeneration efficiency was reduced in old mice, but markedly diminished in geriatric mice (Extended Data Fig. 1f). The number of satellite cells decreased in old versus adult mice, but did not drop further in geriatric mice (Fig. 1a), suggesting that the sharp regenerative decline during sarcopenia cannot be attributed to a reduced satellite cell supply.

Subsequently, we evaluated whether satellite-cell-intrinsic alterations may cause the regenerative decline in geriatric mice. First, muscles of wild-type mice of distinct ages (and SAMP8 mice) were transplanted into young recipient mice (3 months old). One week after transplantation, the number and size of wild-type geriatric and SAMP8 satellite-cell-derived new myofibres, and myonuclei number, were reduced compared to those from adult/old donor wild-type cells or SAMR1 cells, respectively (Fig. 1b, c and Extended Data Fig. 1g). Second, equal numbers of freshly fluorescence-activated cell sorting (FACS)-purified satellite cells from adult, old and geriatric mice (or progeric mice), labelled with green fluorescent protein (GFP), were transplanted into injured muscles of young mice. The number of regenerating GFP$^+$ myofibres derived from geriatric (and progeric) satellite cells was lower than that derived from adult/old cells at 7 and 21 days after transplantation (Fig. 1d and Extended Data Fig. 1h, i). These results indicate that geriatric age induces intrinsic alterations in muscle stem-cell regenerative functions, which cannot be rejuvenated by a young host environment.

[1]Cell Biology Group, Department of Experimental and Health Sciences, Pompeu Fabra University, CIBER on Neurodegenerative diseases, E-08003 Barcelona, Spain. [2]Chromatin and Disease Group, Cancer Epigenetics and Biology Programme, Bellvitge Biomedical Research Institute, L'Hospitalet de Llobregat, E-08907 Barcelona, Spain. [3]Stem Cell Aging Group, Centro Nacional de Investigaciones Cardiovasculares, E-28029 Madrid, Spain. [4]Institució Catalana de Recerca i Estudis Avançats, E-08010 Barcelona, Spain. †Present address: Buck Institute for Research on Aging, Novato, California 94945, USA. *These authors contributed equally to this work.

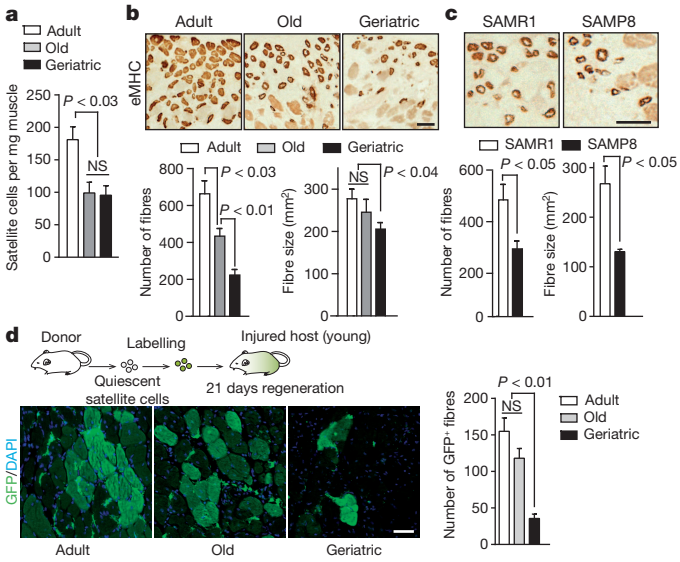


Figure 1 | Intrinsically impaired satellite-cell-dependent muscle regeneration in geriatric and progeric mice. **a**, FACS quantification of $Pax7^+$ satellite cells in resting muscles of wild-type mice at distinct ages: adult (5–6 months), old (20–24 months), geriatric (28–32 months). **b**, **c**, Number and size of $eMHC^+$ fibres in regenerating muscle from wild-type mice (**b**) and SAMR1 and SAMP8 mice (**c**) 1 week after grafting into young mice. Scale bars, 50 µm. **d**, Scheme of transplantation experiment: equal numbers of sorted satellite cells from adult, old and geriatric mice infected with a lentivirus expressing GFP were transplanted into muscles of young immunodeficient mice. After 21 days, muscles were immunostained for GFP. The number of GFP^+ fibres per muscle is shown. Scale bars, 50 µm. Data show mean ± s.e.m. Comparisons by two-sided Mann–Whitney *U*-test. *P* values are indicated. **a**, **b**, $n = 10$ mice; **c**, $n = 8$ donor mice; **d**, $n = 6$ donor mice. **b**, **c**, Two muscles grafted per donor mouse. **d**, At least three independent engraftments per donor mouse. NS, not significant.

Quiescent geriatric satellite cells become pre-senescent

Satellite cells in resting muscle are arrested at G_0 in a reversible fashion (quiescence), maintaining their capacity to activate rapidly in response to injury[13,14]. $Pax7^+$ quiescent satellite cells from geriatric mice showed a reduced activation rate early after injury compared to adult/old cells (Extended Data Fig. 2a). Furthermore, purified quiescent geriatric (and progeric) satellite cells exhibited a defective activation capacity 24 h after transplantation into a young host (Fig. 2a, c), which persisted after an adaptation period before muscle re-injury (Extended Data Fig. 2b) and even on serial transplantations (Fig. 2b), demonstrating the intrinsic nature of the defective geriatric satellite cell activation, and excluding the aged environment as a driving cause for this defect.

To identify the intrinsic factors responsible for the loss of satellite-cell quiescence during sarcopenia, we analysed the global gene expression changes occurring within satellite cells in homeostatic conditions during physiological ageing. Despite differences between young and adult or old satellite cells, geriatric cells have a more divergent gene expression program (Extended Data Fig. 2c and Supplementary Table 1). Through K-means clustering analysis, we isolated a group of genes with increased expression in geriatric satellite cells (cluster G1) (Extended Data Fig. 2d and Supplementary Table 2); cluster G1 was enriched in genes regulated by polycomb repressive complexes (PRC1 and PRC2) (Supplementary Table 2), known mediators of cellular senescence[15]. Comparison of the G1 cluster with a list of senescence-associated genes[16–18] yielded 13 common genes (Supplementary Table 3). Notably, the only significantly upregulated senescence gene in cluster G1 was the master regulator of senescence *p16*^{INK4a} (*Cdkn2a*)[19–22] (Fig. 2d). Inclusion of the progeric satellite cell expression profiling reduced the number of clustered genes (cluster G2), providing a bona fide sarcopenia-associated satellite cell signature (Extended Data Fig. 2e and Supplementary Table 2). These results were surprising because senescence has not been reported

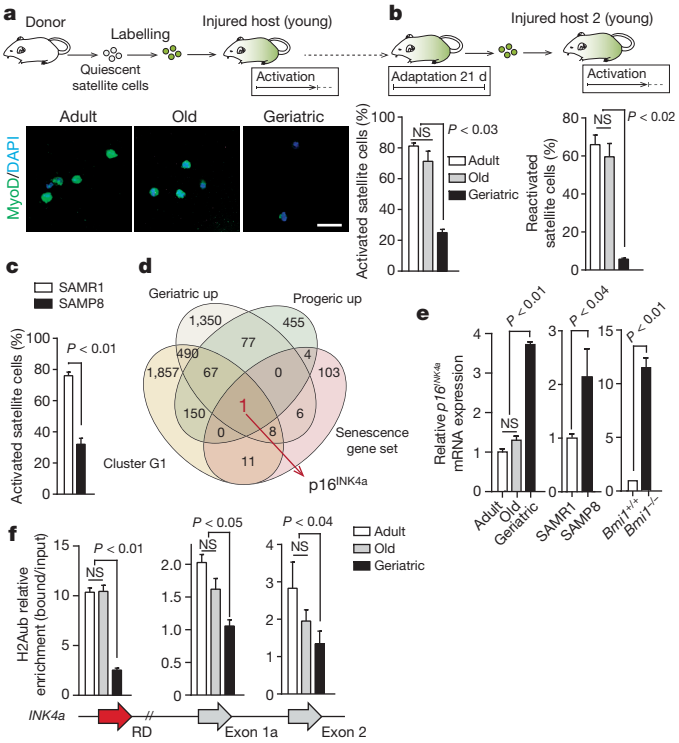


Figure 2 | Satellite cell reversible quiescence is impaired at geriatric age and in progeria. **a**, **b**, Quiescent satellite cells were transplanted as in Fig. 1d after labelling with PKH26 dye (**a**) or GFP-lentivirus (**b**). Cells were sorted 24 h after transplant (**a**) or adapted for 21 days, FACS re-isolated and re-transplanted into a new young host, and reactivation assessed 24 h later (**b**). Quantification of $MyoD^+$ cells from panel **a**. Scale bar, 50 µm. **c**, SAMP8/SAMR1 satellite cells as in panel **a**. **d**, Venn diagram of the overlap among significantly upregulated genes in adult, old and geriatric compared to young satellite cells and progeric SAMP8 compared to SAMR1 satellite cells (false discovery rate (FDR) <0.05 and fold change ≥1.5), genes composing cluster G1, and genes belonging to our senescence gene set (see Extended Data Fig. 2c–g and Supplementary Table 3). **e**, $p16^{INK4a}$ levels. **f**, ChIP analysis for H2AK119ub (H2Aub) in sorted satellite cells from resting muscle. Data are mean ± s.e.m. Two-sided Mann–Whitney *U*-test. **a**–**c**, $n = 5$ donor mice; **e**, **f**, $n = 4$ mice. **a**–**c**, At least three independent engraftments per donor mouse. NS, not significant.

in ageing satellite cells, and no senescence-related genes have been profiled in stem-cell quiescence signatures[23–25].

We confirmed that $p16^{INK4a}$ is induced in resting geriatric and progeric satellite cells (Fig. 2e). Because the canonical PRC1 complex—containing Ring1B E3 ubiquitin ligase and Bmi1—is the major repressor of the *INK4a* locus[26–28], we examined whether H2A monoubiquitination at lysine 119 (H2Aub)[29,30], the PRC1-mediated repressive mark, could distinguish this locus in young, old and geriatric satellite cells. Indeed, H2Aub mark in the *INK4a* locus was significantly reduced in quiescent satellite cells from geriatric mice (Fig. 2f). Interestingly, in adult and old satellite cells, a stronger H2Aub association was observed at the RD domain, which has been demonstrated to be the main binding site for Bmi1 in this locus[31] (Fig. 2f). Thus, loss of PRC1/H2Aub repressive function in geriatric satellite cells results in derepression of the *INK4a* locus and $p16^{INK4a}$ transcriptional induction in homeostatic conditions.

To investigate the effect of anticipated $p16^{INK4a}$ expression in young satellite cells, we used Bmi1-deficient ($Bmi1^{-/-}$) mice, which show premature ageing and muscle wasting[26,32]. Inclusion of Bmi1-deficient satellite cell gene expression profiling into our global clustering analysis (cluster G3) (Extended Data Fig. 2f, g and Supplementary Table 2) confirmed *p16*^{INK4a} as the only senescence-associated gene significantly upregulated in every comparison (Extended Data Fig. 2g), correlating with impaired activation and regenerative capacity of *Bmi1*-null satellite cells (Extended Data Fig. 2h, i).

Thus, geriatric and progeric (and *Bmi1*-null) satellite cells cannot maintain a bona fide reversible quiescence state in homeostatic conditions, and switch to a pre-senescence state, which impairs their further activation.

p16^{INK4a} silencing restores satellite cell quiescence

To demonstrate that p16^{INK4a} is causally implicated in the loss of reversible quiescence of geriatric satellite cells, we assessed the activation capacity of geriatric satellite cells after genetic silencing of *p16*^{INK4a}. Delivery of short hairpin RNA (shRNA) targeting *p16*^{INK4a} (*p16*^{INK4a} shRNA), but not control scrambled shRNA (shScramble), into quiescent geriatric satellite cells downregulated p16^{INK4a} expression and significantly restored activation early after injury, while reducing the expression of senescence-associated genes (Extended Data Fig. 3a). *p16*^{INK4a} silencing also rescued the permanent cell-cycle arrest and allowed geriatric satellite cell activation in culture (Fig. 3a), whereas ectopic p16^{INK4a} overexpression in quiescent young satellite cells prevented their activation (Fig. 3b). *p16*^{INK4a} silencing in freshly sorted geriatric (and progeric and *Bmi1*$^{-/-}$) satellite cells before engraftment into young mice sufficed to restore activation from G_0 arrest and to reduce senescence-associated gene expression (Fig. 3c and Extended Data Fig. 3b, c), and, importantly, this effect persisted even after a 21-day adaptation period before a second injury-induced reactivation (Fig. 3d).

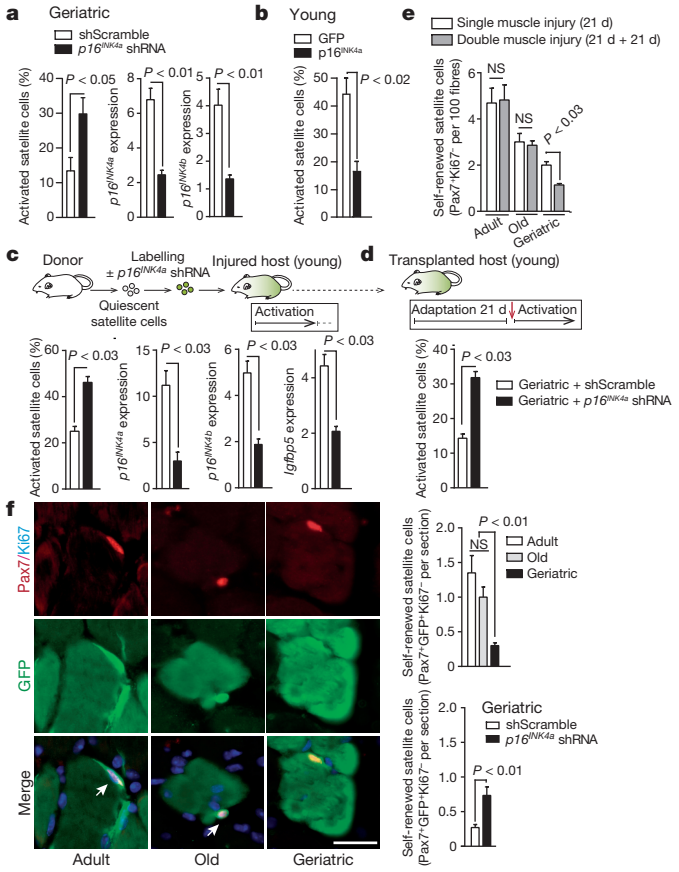


Figure 3 | *p16*^{INK4a} silencing restores reversible quiescence in geriatric satellite cells. **a**, FACS-isolated geriatric resting satellite cells in growth medium, after adenovirus-mediated *p16*^{INK4a} shRNA or shScramble transduction. Activation (BrdU$^+$) and mRNA (quantitative real-time PCR (RT–qPCR)) analyses. **b**, Activation rate (as in Fig. 2a) after p16^{INK4a} or GFP overexpression in young satellite cells. **c**, **d**, *p16*^{INK4a}-silenced geriatric satellite cells were labelled/transplanted as in Fig. 2a (**c**) or Fig. 2b, 21-day-adapted into young host (**d**); activation 24 h after re-transplantation, as in Fig. 2a, and RT–qPCR (**c**). **e**, Self-renewed satellite cells (Pax7$^+$Ki67$^-$) in single and double CTX-injured muscles (at days 21 and 21 + 21, respectively). **f**, 21-day self-renewed satellite cells from **e** with/without *p16*^{INK4a} silencing. Arrows indicate GFP$^+$Pax7$^+$Ki67$^-$ cells. Scale bar, 25 μm. Data show mean ± s.e.m. Two-sided Mann–Whitney *U*-test. **a**, **b**, **e**, $n = 4$ mice; **c**, **d**, $n = 4$ donor mice; **f**, $n = 5$ donor mice; **c**, **d**, **f**, at least three independent engraftments per donor mouse. NS, not significant.

Return to quiescence (self-renewal) is critical to reconstitute the stem-cell pool after tissue damage. To test whether the quiescence-to-senescence switch affects the self-renewal capacity of satellite cells, we performed successive rounds of muscle injury. The presence of self-renewing (Pax7$^+$Ki67$^-$) satellite cells was reduced in geriatric mice 21 days after CTX injury, and this reduction was exacerbated after a second regeneration round (Fig. 3e, f). This impaired self-renewal capacity was confirmed in transplantation experiments of purified geriatric satellite cells, whereas *p16*^{INK4a} silencing rescued this defect (Fig. 3f). The number of self-renewed satellite cells in culture (reserve cells) (Extended Data Fig. 3d) and their activation capacity were also reduced after successive myogenesis rounds, but could be restored by p16^{INK4a} downregulation (Extended Data Fig. 3d,e), whereas p16^{INK4a} overexpression reduced the capacity of young 'reserve' satellite cells to reactivate (Extended Data Fig. 3f). Thus, p16^{INK4a} expression inhibits satellite cell activation and self-renewal in tissue-injury conditions, causing stem-cell depletion over time in sarcopenic muscle.

p16^{INK4a}/Rb/E2F axis induces stem-cell geroconversion

The satellite cell impairment to exit quiescence could also affect the subsequent proliferative expansion required to form new fibres after injury. Indeed, transplanted geriatric satellite cells displayed a strong proliferative impairment during the regeneration process (Fig. 4a) and growth conditions *in vitro* (Extended Data Fig. 4a). Under proliferative pressure, geriatric satellite cells underwent an accelerated entry into deep senescence (that is, geroconversion[33]), as demonstrated by: (1) increased γH2AX DNA repair signal (Fig. 4b and Extended Data Fig. 4b–d); (2) positive immunostaining for senescence-associated proteins p16^{INK4a} and Igfbp5 (Fig. 4b); and (3) positive senescence-associated β-galactosidase (SA-β-gal) staining, a classical senescence marker (Fig. 4c). Transplanted geriatric satellite cells also displayed signs of deep senescence at times of maximal proliferative expansion after injury, as shown by increased expression of γH2AX, p16^{INK4a}, p15^{INK4b}, p19^{ARF}, interleukin (IL)-6 and Igfbp5 (Fig. 4d and Extended Data Fig. 4e), which persisted after a post-transplantation adaptation period within the young host (Fig. 4e–g).

How could proliferating satellite cells become sensitive to increased levels of p16^{INK4a} signalling during ageing? The retinoblastoma (Rb) protein functions downstream of p16^{INK4a} as a co-repressor of E2F proliferation-promoting transcription factors, leading to G_0 irreversible cell-cycle arrest[34,35]. Geriatric satellite cells exposed to proliferative conditions presented high p16^{INK4a} expression levels correlating with reduced phosphorylated Rb protein (Fig. 5a) and low levels of Rb/E2F-regulated genes such as cyclin A, cyclin E, *Mcm3*, *Lmnb1* and *Cdc6* (Fig. 5b). p16^{INK4a} interference reduced geroconversion in geriatric, progeric and *Bmi1*-null satellite cells, and restored muscle regeneration (Fig. 5c–f and Extended Data Fig. 5a–e). Reinforcing these findings, similar results were obtained in *p16*^{INK4a}/*Arf*$^{fl/fl}$ mice[36] after conditional deletion of the *p16*^{INK4a} gene in satellite cells by adenoviral Cre-recombinase transduction (Extended Data Fig. 5f). Conversely, forced p16^{INK4a} expression promoted geroconversion of young satellite cells during regeneration and *in vitro* myogenesis (Extended Data Fig. 5g, h). Supporting the p16^{INK4a}/Rb-driven geroconversion of satellite cells, we found a substantial reduction in the expression of Rb/E2F-regulated genes in geriatric satellite cells transplanted into young host muscle under regenerative/proliferative pressure (Extended Data Fig. 5i), coinciding with increased expression of senescence-associated genes (Fig. 4d), and both effects were reverted by p16^{INK4a} interference (Fig. 5d, e).

Active p16^{INK4a}/Rb axis in human geriatric stem cells

As in murine satellite cells, p16^{INK4a} was exclusively expressed in human satellite cells from geriatric individuals (of about 75 years) compared to

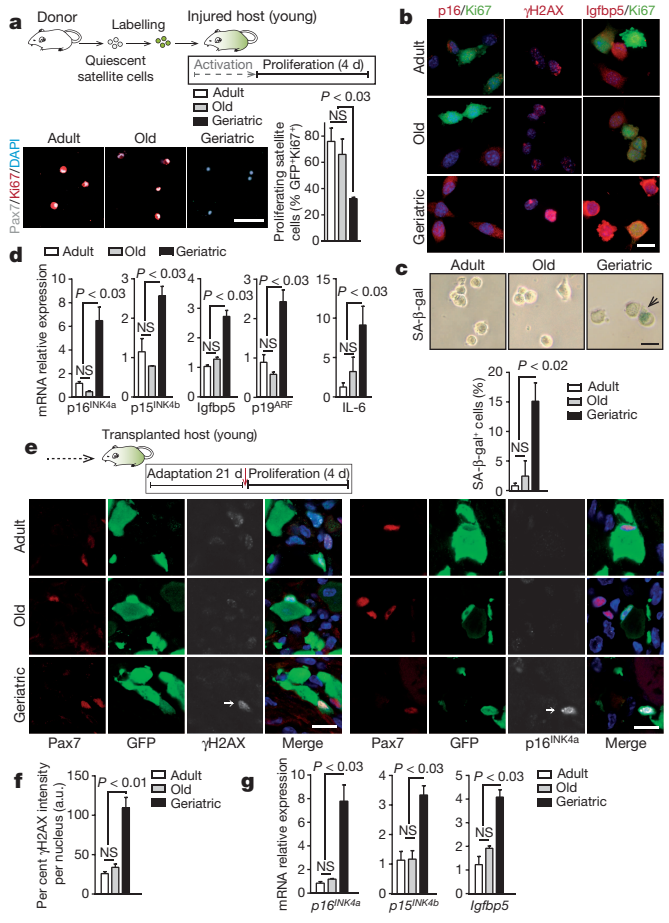


Figure 4 | Satellite cell geroconversion is caused by p16^{INK4a}. **a**, GFP$^+$Ki67$^+$ satellite cells quantified after transplantation (as in Fig. 1d) 4 days after CTX injury. Scale bar, 50 μm. **b**, **c**, Sorted satellite cells cultured in growth medium and stained for p16^{INK4a}, γH2AX, Igfbp5 and Ki67 (DAPI) (**b**) or SA-β-gal (**c**). **d**, qRT–PCR in cells from panel **a**. **e**, Transplanted satellite cells from panel **a** adapted 21 days in young host; GFP$^+$ satellite cells were sorted 4 days after re-injury and Pax7 and γH2AX analysed. Scale bars, 50 μm. **f**, γH2AX levels per nucleus on Pax7$^+$ cells from panel **e**. **g**, qRT–PCR in cells from panel **e**. Data are mean ± s.e.m. Two-sided Mann–Whitney *U*-test. **a**, **d**–**g**, $n = 4$ donor mice; **b**, **c**, $n = 5$ biological replicates. **a**, **d**–**g**, At least three independent engraftments per donor mouse. NS, not significant.

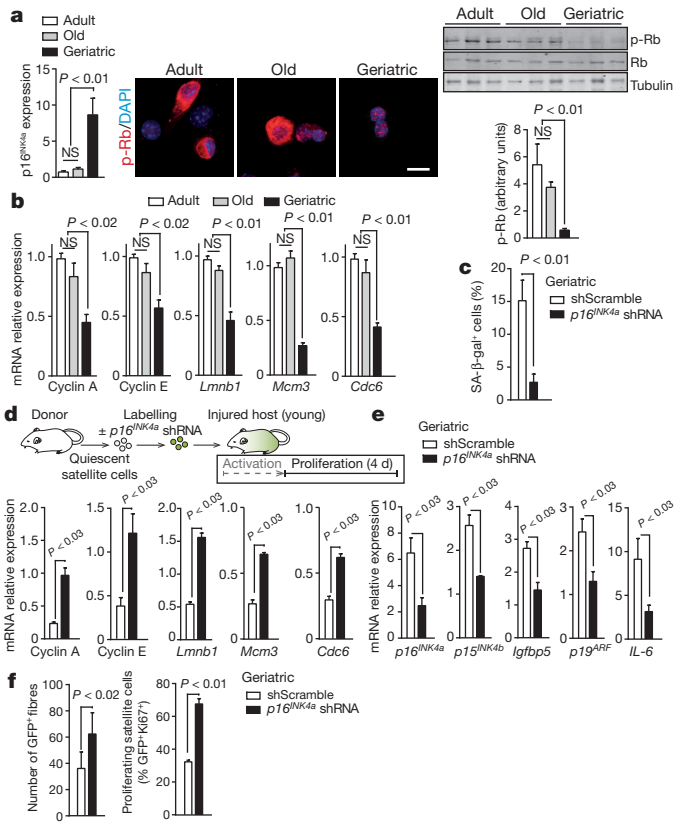


Figure 5 | p16^{INK4a}-driven Rb/E2F axis regulates geroconversion. **a**, Sorted satellite cells cultured in growth medium: p16^{INK4a} quantification; phosphorylated Rb (p-Rb) immunostaining/DAPI (scale bar, 10 μm); p-Rb/total Rb western blotting. **b**, RT–qPCR of Rb/E2F targets in cells from panel **a**. **c**, Geriatric cells from panel **a**, transduced with adenovirus-mediated *p16*^{INK4a} shRNA/shScramble, cultured in growth medium. Quantification of SA-β-gal$^+$ cells. **d**, **e**, Geriatric satellite cells (transduced with *p16*^{INK4a} shRNA or shScramble lentiviruses) transplanted as in Fig. 4a, re-sorted and analysed after 4 days. **f**, Proliferating GFP$^+$ satellite cells and GFP$^+$ fibres in 4-day transplanted muscles from panel **d**. Data are mean ± s.e.m. Two-sided Mann–Whitney *U*-test. **a**, $n = 5$ biological replicates; **b**, **c**, $n = 6$ biological replicates; **d**–**f**, $n = 4$ donor mice. **d**–**f**, At least three independent engraftments per donor mouse. NS, not significant.

young counterparts (of about 25 years) (Fig. 6a), and this correlated with signs of sarcopenia in muscle biopsies of geriatric individuals (Extended Data Fig. 6a). Consistent with this finding, satellite cells expressed p16^{INK4a} and Igfbp5 in biopsies of sarcopenic muscle from geriatric subjects, but not from young subjects (Extended Data Fig. 6b). Furthermore, p16^{INK4a} expression in human geriatric satellite cells prevented myogenic functions (Extended Data Fig. 6c, d), while inducing senescence (Fig. 6a and 6b). These defects were probably caused by dysregulation of the Rb–E2F pathway, as demonstrated by the reduction of phosphorylated Rb and E2F target-gene expression (Fig. 6c, d). Genetic interference with human p16^{INK4a} restored geriatric satellite cell proliferation by reducing senescence (Fig. 6e). Thus, p16^{INK4a} expression is specifically induced in geriatric satellite cells and drives geroconversion at the expense of proliferation both in humans and mice.

Discussion

Reversible quiescence is a defining property of adult mammalian stem cells, especially in relatively stable tissues with low turnover[13]. Our results demonstrate that the molecular regulation of quiescence in adult life depends on the active repression of genes associated with senescence and is integrated in the transcriptional and epigenetic network that regulates distinct fates of stem-cell progeny during ageing. We provide evidence of an abnormal satellite cell fate in sarcopenic muscle of geriatric and progeric mice, by impaired maintenance of the homeostatic quiescence state, which instead switches to a pre-senescence state, resulting in inhibition of the regenerative capacity of muscle. This quiescence-to-senescence cellular shift was surprising considering that senescence of muscle stem cells has not been reported previously.

Mechanistically, we show that in muscle homeostatic conditions, adult or even old satellite cells preserve their reversible quiescent state through repression of p16^{INK4a}. In older (geriatric) satellite cells, loss of bona fide reversible quiescence is caused by derepression of the *p16*^{INK4a} locus, through reduction of H2A ubiquitination, a critical epigenetic mark in PRC1 gene repression[29,30]. Consistent with this notion, loss of the essential PRC1 member *Bmi1* in young satellite cells led to derepression of the locus and p16^{INK4a} upregulation, inducing an irreversible senescent-like fate and preventing satellite-cell-mediated regeneration. We speculate that this derepression is a cell-autonomous response to increasing accumulation of genomic damage with ageing. Unless exposed to injury or in degenerative conditions, muscle stem cells have a low turnover rate throughout life, which potentially allows for the accumulation of genomic damage during the homeostatic state of quiescence, just like in post-mitotic somatic cells, which could promote acquisition of a pre-senescence-like state[10].

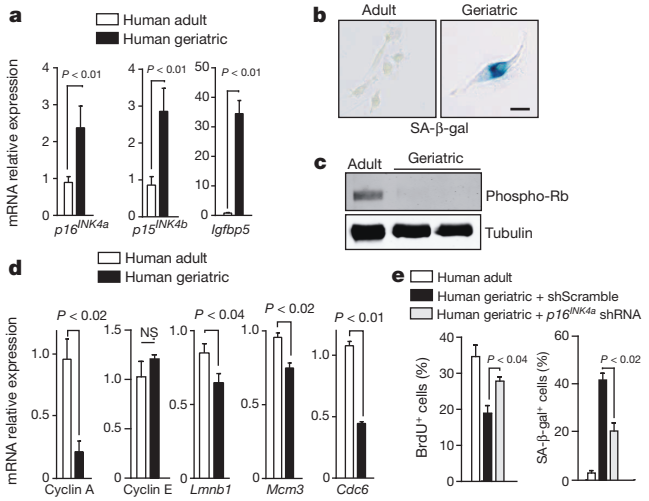


Figure 6 | p16^{INK4a}/Rb/E2F senescence pathway in human geriatric satellite cells. **a**, RT–qPCR of *p16^{INK4a}*, *p15^{INK4b}* and *Igfbp5* in proliferating human satellite cells from adult (25 ± 4 years) or geriatric (75 ± 4 years) donors. **b**, Representative pictures of the cells described in panel **a** assayed for SA-β-gal activity. **c**, **d**, Phosphorylated Rb western blotting and RT–qPCR of Rb/E2F targets in cells from panel **a**. **e**, *p16^{INK4a}* silencing in cells from panel **a** in growth medium: BrdU^{+} and SA-β-gal^{+} cells. Scale bar, 10 μm. Data are mean ± s.e.m. Two-sided Mann–Whitney *U*-test. **a**–**e**, $n = 5$ biological replicates. NS, not significant.

Recent studies showed that production of fibroblast growth factor-2 by the ageing muscle fibre brakes quiescence of old satellite cells by inducing proliferation[37]. We did not observe signs of proliferation nor apoptosis in geriatric or progeric satellite cells in homeostatic conditions (not shown). Instead, we show that geriatric satellite cells lose the normal quiescent state through a quiescence-to-senescence switch, which stalls cells in an irreversible cell-cycle arrest. When exposed to regenerative pressure, these pre-senescent satellite cells, in an attempt to proliferate in the continuous presence of growth factors, while the p16^{INK4a}-induced cell-cycle arrest persists, acquire a full senescence state (geroconversion), which leads to impaired stem-cell self-renewal and depletion of the satellite cell pool over time. Because this phenomenon also occurs in mice with diseases that recapitulate some aspects of ageing, such as SAMP8 progeria mice, the loss of quiescence through the p16^{INK4a}-dependent pathway might be a general feature of stem cells at advanced age. Consistent with this, interference with senescence was shown to attenuate premature organ ageing in mice with BubR1 insufficiency[22,38,39].

Our study also uncovered that, in geriatric satellite cells, p16^{INK4a} inhibition of Rb drives accelerated senescence, based on the known capacity of the p16^{INK4a}–Rb pathway to form senescence-associated heterochromatic foci and accumulate DNA-repair foci marked by γH2AX (reviewed in refs 15, 40, 41), at the expense of the normal myogenic fate. Interestingly, Smad–Notch signalling imbalance was shown to induce the expression of p15^{INK4b} and p21$^{CIP1/WAF1}$ CDK inhibitors in proliferating satellite cells of old (non-geriatric) mice, although p16^{INK4a} expression was undetectable[42]. We therefore propose that acquisition of p16^{INK4a} by satellite cells with age constitutes a point of 'no return' in their capacity to maintain quiescence and regenerative functions. The relevance of our data is further supported by the finding that p16^{INK4a} also appeared to be positively associated with reduced myogenic potential and increased cellular senescence in human satellite cells from geriatric individuals with sarcopenia.

In conclusion, our studies demonstrate that maintenance of the reversible quiescent state in resting muscle depends on the active repression of senescence pathways, and provide novel evidence for age-related intrinsic changes in satellite cells, accounting for loss of homeostasis and defective regeneration, which cannot be rejuvenated by a youthful environment. However, we propose that senescence is not an insurmountable barrier to regeneration, as p16^{INK4a} inhibition sufficed to rejuvenate satellite cells and restore regeneration in physiologically aged and progeric mice. This offers what we believe to be a novel strategy for attenuating defective stem-cell functions in patients with progeroid syndromes and potentially in old age with severely affected tissue regenerative capacity.

METHODS SUMMARY

Male mice (C57BL/6 (wild type), *Bmi1$^{-/-}$*, *p16$^{INK4A/}$Arf$^{fl/fl}$* and SAMP8 and SAMR1) were used at different ages. Muscle damage/regeneration was induced by cardiotoxin injection or grafting. Satellite cells of distinct ages/genotypes were FACS isolated and transplanted into young hosts (with/without *p16^{INK4a}* silencing), and satellite cell activation, proliferation, myofibre formation, self-renewal or senescence were analysed subsequently.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 June 2013; accepted 10 January 2014.
Published online 12 February; corrected online 19 February 2014 (see full-text online HTML version for details).

1. Burtner, C. R. & Kennedy, B. K. Progeria syndromes and ageing: what is the connection? *Nature Rev. Mol. Cell Biol.* **11,** 567–578 (2010).
2. Arthur, S. T. & Cooley, I. D. The effect of physiological stimuli on sarcopenia; impact of Notch and Wnt signaling on impaired aged skeletal muscle repair. *Int. J. Biol. Sci.* **8,** 731–760 (2012).
3. Jang, Y. C., Sinha, M., Cerletti, M., Dall'Osso, C. & Wagers, A. J. Skeletal muscle stem cells: effects of aging and metabolism on muscle regenerative function. *Cold Spring Harb. Symp. Quant. Biol.* **76,** 101–111 (2011).
4. Renault, V., Thornell, L. E., Eriksson, P. O., Butler-Browne, G. & Mouly, V. Regenerative potential of human skeletal muscle during aging. *Aging Cell* **1,** 132–139 (2002).
5. Cheung, T. H. & Rando, T. A. Molecular regulation of stem cell quiescence. *Nature Rev. Mol. Cell Biol.* **14,** 329–340 (2013).
6. Yin, H., Price, F. & Rudnicki, M. A. Satellite cells and the muscle stem cell niche. *Physiol. Rev.* **93,** 23–67 (2013).
7. Shefer, G., Van de Mark, D. P., Richardson, J. B. & Yablonka-Reuveni, Z. Satellite-cell pool size does matter: defining the myogenic potency of aging skeletal muscle. *Developmental biology* **294,** 50–66 (2006).
8. Carlson, M. E. & Conboy, I. M. Loss of stem cell regenerative capacity within aged niches. *Aging Cell* **6,** 371–382 (2007).
9. García-Prat, L., Sousa-Victor, P. & Munoz-Canoves, P. Functional dysregulation of stem cells during aging: a focus on skeletal muscle stem cells. *FEBS J.* **280,** 4051–4062 (2013).
10. Liu, L. & Rando, T. A. Manifestations and mechanisms of stem cell aging. *J. Cell Biol.* **193,** 257–266 (2011).
11. Smythe, G. M. *et al.* Age influences the early events of skeletal muscle regeneration: studies of whole muscle grafts transplanted between young (8 weeks) and old (13–21 months) mice. *Exp. Gerontol.* **43,** 550–562 (2008).
12. Derave, W., Eijnde, B. O., Ramaekers, M. & Hespel, P. No effects of lifelong creatine supplementation on sarcopenia in senescence-accelerated mice (SAMP8). *Am. J. Physiol. Endocrinol. Metab.* **289,** E272–E277 (2005).
13. Dhawan, J. & Rando, T. A. Stem cells in postnatal myogenesis: molecular mechanisms of satellite cell quiescence, activation and replenishment. *Trends Cell Biol.* **15,** 666–673 (2005).
14. Shea, K. L. *et al.* Sprouty1 regulates reversible quiescence of a self-renewing adult muscle stem cell pool during regeneration. *Cell Stem Cell* **6,** 117–129 (2010).
15. Lanigan, F., Geraghty, J. G. & Bracken, A. P. Transcriptional regulation of cellular senescence. *Oncogene* **30,** 2901–2911 (2011).
16. Fridman, A. L. & Tainsky, M. A. Critical pathways in cellular senescence and immortalization revealed by gene expression profiling. *Oncogene* **27,** 5975–5987 (2008).
17. Collado, M. *et al.* Tumour biology: senescence in premalignant tumours. *Nature* **436,** 642 (2005).
18. Coppé, J. P., Desprez, P. Y., Krtolica, A. & Campisi, J. The senescence-associated secretory phenotype: the dark side of tumor suppression. *Annu. Rev. Pathol.* **5,** 99–118 (2010).
19. Gonzalez, S. *et al.* Oncogenic activity of Cdc6 through repression of the *INK4/ARF* locus. *Nature* **440,** 702–706 (2006).
20. Janzen, V. *et al.* Stem-cell ageing modified by the cyclin-dependent kinase inhibitor p16^{INK4a}. *Nature* **443,** 421–426 (2006).
21. Molofsky, A. V. *et al.* Increasing *p16^{INK4a}* expression decreases forebrain progenitors and neurogenesis during ageing. *Nature* **443,** 448–452 (2006).
22. Baker, D. J. *et al.* Clearance of p16^{Ink4a}-positive senescent cells delays ageing-associated disorders. *Nature* **479,** 232–236 (2011).
23. Fukada, S. *et al.* Molecular signature of quiescent satellite cells in adult skeletal muscle. *Stem Cells* **25,** 2448–2459 (2007).
24. Liu, L. *et al.* Chromatin modifications as determinants of muscle stem cell quiescence and chronological aging. *Cell Rep.* **4,** 189–204 (2013).

25. Pallafacchina, G. *et al.* An adult tissue-specific stem cell in its niche: a gene profiling analysis of *in vivo* quiescent and activated muscle satellite cells. *Stem Cell Res.* **4,** 77–91 (2010).
26. Jacobs, J. J., Kieboom, K., Marino, S., DePinho, R. A. & van Lohuizen, M. The oncogene and Polycomb-group gene *bmi-1* regulates cell proliferation and senescence through the *ink4a* locus. *Nature* **397,** 164–168 (1999).
27. Bracken, A. P. *et al.* The Polycomb group proteins bind throughout the INK4A-ARF locus and are disassociated in senescent cells. *Genes Dev.* **21,** 525–530 (2007).
28. Margueron, R. *et al.* Ezh1 and Ezh2 maintain repressive chromatin through different mechanisms. *Mol. Cell* **32,** 503–518 (2008).
29. Wang, H. *et al.* Role of histone H2A ubiquitination in Polycomb silencing. *Nature* **431,** 873–878 (2004).
30. Cao, R., Tsukada, Y. & Zhang, Y. Role of Bmi-1 and Ring1A in H2A ubiquitylation and Hox gene silencing. *Mol. Cell* **20,** 845–854 (2005).
31. Agherbi, H. *et al.* Polycomb mediated epigenetic silencing and replication timing at the *INK4a/ARF* locus during senescence. *PLoS ONE* **4,** e5622 (2009).
32. Robson, L. G. *et al.* Bmi1 is expressed in postnatal myogenic satellite cells, controls their maintenance and plays an essential role in repeated muscle regeneration. *PLoS ONE* **6,** e27116 (2011).
33. Blagosklonny, M. V. Selective anti-cancer agents as anti-aging drugs. *Cancer Biol. Ther.* **14,** 1092–1097 (2013).
34. Chicas, A. *et al.* Dissecting the unique role of the retinoblastoma tumor suppressor during cellular senescence. *Cancer Cell* **17,** 376–387 (2010).
35. Trimarchi, J. M. & Lees, J. A. Sibling rivalry in the E2F family. *Nature Rev. Mol. Cell Biol.* **3,** 11–20 (2002).
36. Krimpenfort, P., Quon, K. C., Mooi, W. J., Loonstra, A. & Berns, A. Loss of $p16^{Ink4a}$ confers susceptibility to metastatic melanoma in mice. *Nature* **413,** 83–86 (2001).
37. Chakkalakal, J. V., Jones, K. M., Basson, M. A. & Brack, A. S. The aged niche disrupts muscle stem cell quiescence. *Nature* **490,** 355–360 (2012).
38. Baker, D. J. *et al.* Opposing roles for $p16^{Ink4a}$ and $p19^{Arf}$ in senescence and ageing caused by BubR1 insufficiency. *Nature Cell Biol.* **10,** 825–836 (2008).
39. Baker, D. J., Weaver, R. L. & van Deursen, J. M. p21 both attenuates and drives senescence and aging in BubR1 progeroid mice. *Cell Rep.* **3,** 1164–1174 (2013).
40. d'Adda di Fagagna, F. Living on a break: cellular senescence as a DNA-damage response. *Nature Rev. Cancer* **8,** 512–522 (2008).
41. Sperka, T., Wang, J. & Rudolph, K. L. DNA damage checkpoints in stem cells, ageing and cancer. *Nature Rev. Mol. Cell Biol.* **13,** 579–590 (2012).
42. Carlson, M. E., Hsu, M. & Conboy, I. M. Imbalance between pSmad3 and Notch induces CDK inhibitors in old muscle stem cells. *Nature* **454,** 528–532 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are indebted to M. Raya and V. Lukesova for their contributions to this study; J. Martín-Caballero (PRBB Animal Facility), O. Fornas (UPF/CRG FACS Facility) and CRG/UPF Genomic Facility for technical help; A. Consiglio for help in lentivirus obtention; E. Rebollo for advice on imaging; M. van Lohuizen for Bmi1-deficient mice; M. Blasco's laboratory for help with ageing mice; T. Kawamura and M. Serrano for $p16^{INK4a}$ constructs; A. Sacco, S. Tajbakhsh, B. Gayraud-Morel, D. Montarras, J. Morgan and F. S. Tedesco for advice on cell transplantation; A. Brack and P. Zammit for advice on reserve cells; and Myoage network and tissue bank for support. The authors acknowledge funding from MINECO-Spain (SAF2012-38547, FIS-PS09/01267, FIS-PI13/025, PLE2009-0124), AFM, MDA, E-Rare, Fundació Marató TV3, DuchennePP-NL and EU-FP7 (Myoage, Optistem and Endostem). P.S.-V. and L.G.-P. were supported by predoctoral fellowships from Fundação para a Ciência e a Tecnologia (Portugal) and Programa de Formación de Personal Investigador (Spain), respectively.

Author Contributions P.S.-V. designed and performed experiments, analysed data, interpreted results and wrote the manuscript. S.Gu. and L.G.-P. designed and performed experiments, analysed data and interpreted results. L.O., V.R.-B. and M.J. performed experiments and provided technical support. J.R.-U. and E.B. performed ChIP experiments and edited the manuscript. S.Go. generated transgenic mice and edited the manuscript. A.L.S. and E.P. conceived the project, designed and performed experiments, interpreted results and wrote the manuscript. P.M.-C. conceived the project, designed experiments, interpreted results and wrote the manuscript.

Author Information Microarray data have been deposited into the NCBI Gene Expression Omnibus under accession number GSE53728. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.M.-C. (pura.munoz@upf.edu) or E.P. (eusebio.perdiguero@upf.edu).

ARTICLE

doi:10.1038/nature12948

In situ identification of bipotent stem cells in the mammary gland

Anne C. Rios[1,2]*, Nai Yang Fu[1,2]*, Geoffrey J. Lindeman[1,3,4] & Jane E. Visvader[1,2]

The mammary epithelium undergoes profound morphogenetic changes during development. Architecturally, it comprises two primary lineages, the inner luminal and outer myoepithelial cell layers. Two opposing concepts on the nature of mammary stem cells (MaSCs) in the postnatal gland have emerged. One model, based on classical transplantation assays, postulates that bipotent MaSCs have a key role in coordinating ductal epithelial expansion and maintenance in the adult gland, whereas the second model proposes that only unipotent MaSCs identified by lineage tracing contribute to these processes. Through clonal cell-fate mapping studies using a stochastic multicolour *cre* reporter combined with a new three-dimensional imaging strategy, we provide evidence for the existence of bipotent MaSCs as well as distinct long-lived progenitor cells. The cellular dynamics at different developmental stages support a model in which both stem and progenitor cells drive morphogenesis during puberty, whereas bipotent MaSCs coordinate ductal homeostasis and remodelling of the mouse adult gland.

The mammary gland is composed of a branching ductal tree embedded in a stromal matrix. The epithelial tree comprises cells of two lineages, the myoepithelial and luminal, which can be subdivided into ductal and alveolar subtypes. Development of the gland[1] predominantly occurs during puberty, involving extensive branching and elongation of the ducts via terminal end buds (TEBs) to generate a complex ductal network. With each oestrus cycle, the epithelium enters cycles of proliferation and differentiation, leading to the emergence of small alveolar buds. During pregnancy, the alveolar epithelium undergoes massive expansion to yield milk-producing alveoli. The profound regenerative capacity evident on successive rounds of pregnancy has implied the presence of renewable stem cells[2–4]. MaSCs and distinct luminal progenitor cell types have been prospectively isolated from adult mouse and human mammary tissue[5–13], pointing to the existence of an epithelial differentiation hierarchy. In the mouse, there is accumulating evidence for a heterogeneous stem cell compartment comprising long-term and short-term repopulating cells based on transplantation assays[14], as well as for slow-cycling stem cells[15,16] and long-term label-retaining cells capable of asymmetric cell division[17]. A considerable increase in MaSC activity occurs during pregnancy[14] and the dioestrus phase in cycling female mice[18]. Together these data have led to the concept that bipotential MaSCs have a fundamental role in the postnatal mammary gland.

An important alternative approach for assessing stem cell function is lineage tracing, which allows stem and progenitor cell fate to be tracked *in situ* in the context of development, tissue maintenance and disease[19]. Recent lineage-tracing studies in the mammary gland[20,21] have described unipotent stem cells that separately control each lineage throughout development. In one study[20], unipotent luminal and myoepithelial cells were identified in the postnatal mammary gland during puberty, adulthood and pregnancy, leading to the conclusion that only unipotent cells govern homeostasis and the key stages of ontogeny. A subsequent lineage-tracing study added a further layer of complexity through the discovery that Wnt-responsive Axin2$^+$ cells in the mouse mammary gland can switch fate according to the developmental stage[21]. Furthermore, the basal-restricted Axin2$^+$ cells defined *in vivo* behaved as bipotent stem cells when transplanted, suggesting that the regenerative potential unmasked by transplantation may not be physiologically relevant[21].

Collectively, these findings have called into question the existence of bipotent stem cells in the adult mammary gland. Here we address whether such cells exist and the relative contribution of stem versus progenitor cells to mammary development and homeostasis by combining clonal lineage tracing with a novel three-dimensional (3D) imaging strategy using newly generated lineage-specific strains. We provide *in vivo* evidence for an epithelial differentiation hierarchy in which bipotent MaSCs exert key physiological roles in the postnatal mammary gland.

Novel 3D imaging at high cellular resolution

To explore the contribution of stem versus progenitor cells to the developing mammary gland *in situ*, we developed a 3D confocal imaging strategy that allows visualization of expansive areas of the epithelial tree in its native stroma. Ductal architecture could be visualized at high cellular resolution, spanning up to 3 mm of tissue (Fig. 1a and Supplementary Videos 1–3). By contrast, the two-dimensional confocal microscopy routinely used in lineage tracing only covers a limited number of microns. Luminal and myoepithelial cells were readily distinguished by their cuboidal and highly elongated (some spanning 100 microns) shapes, respectively. Moreover, immunostaining using lineage-specific markers could be combined with 3D imaging to validate cell fate at single-cell resolution (Fig. 1b, d).

Generation of inducible lineage-specific strains

We generated reverse tetracycline transcriptional activator (rtTA)-internal ribosome entry site (IRES)-green fluorescent protein (GFP) transgenic strains that were predicted to be specific for the basal and luminal epithelial lineages. To potentially mark committed luminal progenitor cells, we selected the transcription factor *Elf5*, one of the most highly ranked genes in the luminal progenitor gene signature for both mouse and human[22] and a key regulator of alveolar differentiation[23].

[1]Stem Cells and Cancer Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia. [2]Department of Medical Biology, The University of Melbourne, Parkville, Victoria 3010, Australia. [3]Department of Medicine, The University of Melbourne, Parkville, Victoria 3010, Australia. [4]Department of Medical Oncology and Familial Cancer Centre, The Royal Melbourne Hospital, Parkville, Victoria 3050, Australia.
*These authors contributed equally to this work.

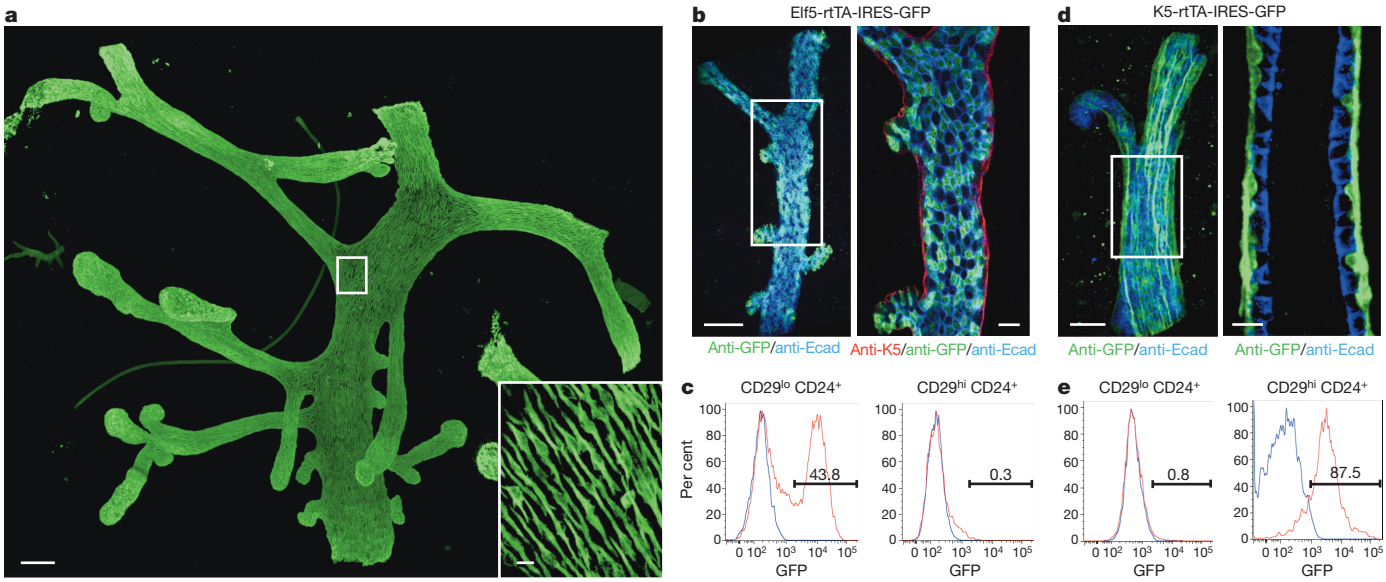


Figure 1 | A novel 3D imaging strategy for validation of lineage-specific reporter strains and cell-fate mapping. **a**, 3D confocal reconstruction of a whole-mounted ductal tree (3 weeks old) immunostained for keratin 5 (K5). This tile scan image represents 90 *Z*-stacks containing 368 optical sections. The enlargement shows a thick optical section (5 μm) at higher magnification. Scale bars, 100 μm (whole-mount) and 10 μm (inset, optical section). **b**–**e**, Validation of the specificity of the *Elf5* promoter for luminal cells in Elf5-rtTA-IRES-GFP mammary glands (**b**, **c**) and of the *K5* promoter for basal cells in K5-rtTA-IRES-GFP glands (**d**, **e**) at 8 weeks of age. **b**, Left, whole-mount 3D confocal image showing GFP (green) expression from the *Elf5* promoter, E-cadherin (Ecad, also known as Cdh1; blue) and K5 expression (red). Right, an optical section (0.45 μm) of the rectangular area. **c**, Representative FACS histograms showing GFP expression in the basal ($Lin^-CD29^{hi}CD24^+$) and luminal ($Lin^-CD29^{lo}CD24^+$) populations of Elf5-rtTA-IRES-GFP (red) and wild-type (blue) mammary glands from FVB/N female mice ($n = 3$ for each). **d**, Left, whole-mount confocal image showing GFP expression (green) from the *K5* promoter and E-cadherin (blue). Right, a thick optical section (3 μm). Scale bars, 50 μm (whole-mount) and 20 μm (inset, optical section). **e**, Representative FACS histograms showing GFP expression in the basal and luminal populations from K5-rtTA-IRES-GFP (red) and wild-type (blue) glands from FVB/N mice ($n = 3$).

Whole-mount confocal imaging of Elf5-rtTA-IRES-GFP transgenic mice showed that *Elf5*-GFP was restricted to the inner luminal layer (Fig. 1b), whereas flow cytometric analysis confirmed that GFP^+ cells were confined to the luminal ($CD29^{lo}$ $CD24^+$) but not the basal/MaSC-enriched population ($CD29^{hi}$ $CD24^+$) in pubertal and adult mammary glands (Fig. 1c and Extended Data Fig. 1a, b). Strong concordance was found between endogenous Elf5 expression and GFP driven from the *Elf5* promoter (Extended Data Fig. 2a–d). Fractionation into luminal progenitor ($CD61^+$) and mature luminal ($CD61^-$) subsets[5] showed that the *Elf5*-GFP^+ population encompassed the $CD61^+$ subset, whereas colony-forming assays revealed that only *Elf5*-GFP^+ cells had clonogenic activity, indicating that all luminal progenitors express *Elf5* (Extended Data Fig. 2).

To direct expression to the basal population, we used the bovine *keratin 5* (*K5*, also known as *Krt5*) promoter, which faithfully recapitulates expression of the endogenous *K5* gene and has been extensively used for *in vivo* studies in the mouse[24,25]. It is important to note that the basal population comprises MaSCs, myoepithelial cells and probably other basal cell intermediates[8,12]. GFP expression was confined to the outer myoepithelial layer by whole-mount confocal microscopy (Fig. 1d), and fluorescence-activated cell sorting (FACS) analysis of K5-rtTA-IRES-GFP mammary glands confirmed that only the basal population contained GFP^+ cells (Fig. 1e and Extended Data Fig. 1c).

Identification of restricted luminal progenitors

To track the *in vivo* fate and spatial distribution of *Elf5*-expressing cells during ductal expansion during puberty, we exploited the multicolour Confetti system, a powerful tool for clonal analysis, as demonstrated for intestinal crypts[26]. To do so, we generated triple-transgenic mice harbouring the Elf5-rtTA-IRES-GFP, TetO-cre and floxed R26R-Confetti reporter alleles[26]. Cre-mediated recombination within the floxed reporter locus then randomly activates one of its four coloured reporters and thus indelibly marks the recombined cell and its progeny for their entire lifespan (Extended Data Fig. 3). Elf5-rtTA/TetO-cre/R26R-Confetti transgenic mice were administered a single dose of doxycycline at the onset of puberty and analysed 2 days later to establish the labelling efficiency and then at 2 weeks post-induction, when morphogenesis of the mammary gland is largely complete. Whole-mount confocal microscopy revealed a sea of multicoloured cells within the TEBs and along the ducts (Fig. 2a, b). Similarly, a high frequency of multicoloured cells was observed along the ducts of young adults at 8 weeks post-induction (Fig. 2c), with equal representation of the four colours (Extended Data Fig. 3d).

To evaluate the kinetics of cell labelling more quantitatively, the single-colour R26R-tdTomato locus was crossed into Elf5-rtTA/TetO-cre mice to allow FACS analysis. Notably, a high labelling efficiency (85–95%) could be achieved with the *Elf5*-driven model. Only luminal cells were labelled through puberty and into adulthood, confirming the exquisite lineage specificity of *Elf5* (Fig. 2d). Within the total luminal population, approximately 25% of tdTomato-positive cells were observed 2 days post-induction, reaching ~40% at 8 weeks. This labelled population remained stable between the end of puberty and adulthood.

Luminal progenitors contribute to ductal maintenance

To address the role of *Elf5*-expressing cells in adulthood, Elf5-rtTA/TetO-cre/R26R-tdTomato mice were induced with doxycycline at 9 weeks of age to distinguish ductal morphogenesis in puberty from the maintenance phase in adulthood. The percentage of tdTomato-positive luminal cells remained unchanged between the 2-day and 8-week chases (Fig. 2e), indicating that *Elf5*-marked luminal progenitors are relatively long-lived, reminiscent of those in the interfollicular epidermis[27]. Confocal imaging of glands from Elf5-rtTA/TetO-cre mice on the R26R-Confetti background confirmed the luminal identity of labelled cells in the ducts. The presence of multicoloured ducts and alveolar buds directly implicates progenitors in ductal homeostasis and bud formation (Fig. 2f). Quantification of clones revealed that most were composed of only a few cells, and this did not change markedly over an 8-week chase (Extended Data Fig. 2e–i). The sparse distribution of

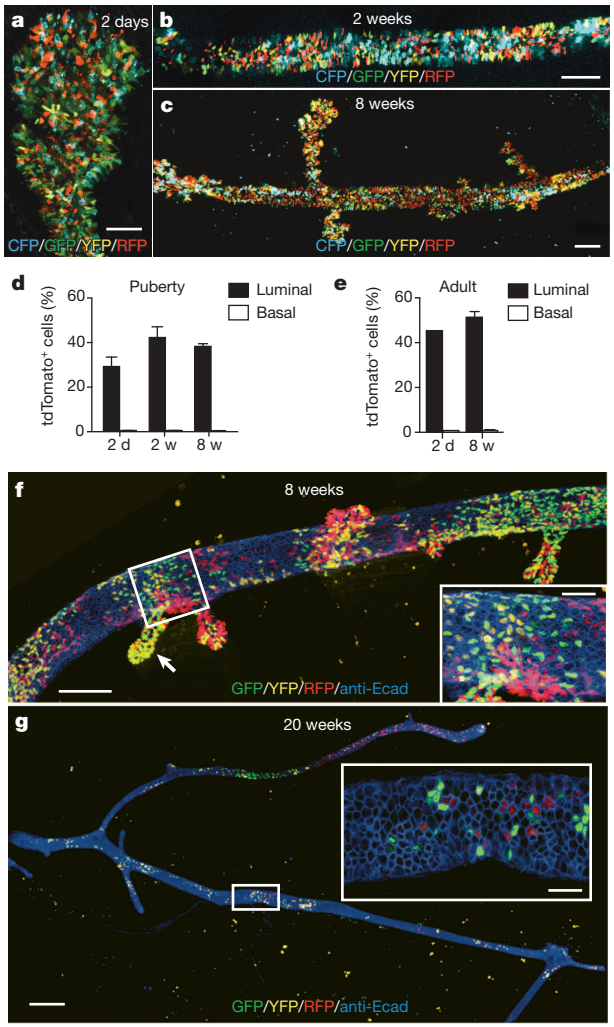


Figure 2 | Elf5-expressing luminal progenitor cells contribute to morphogenesis in puberty and maintenance in adulthood. **a**–**c**, Whole-mount 3D confocal images of a TEB or duct from Elf5-rtTA/TetO-cre/R26R-Confetti mice analysed at 2 days (**a**), 2 weeks (**b**) or 8 weeks (**c**) after doxycycline injection during puberty (4 weeks old). Scale bars, 50 μm (**a**, TEB) and 100 μm (**b**, **c**, ducts). CFP, cyan fluorescence protein. **d**, **e**, Bar charts showing the percentage of tdTomato-expressing cells in the luminal and basal populations from Elf5-rtTA/TetO-cre/R26R-tdTomato mice (mixed background) at the indicated times post-induction in puberty (**d**) and adulthood (**e**). d, days; w, weeks. For **d**, **e**, data represent the mean ± s.e.m. for $n = 3$ mice. **f**, **g**, Whole-mount 3D confocal images of ducts from Elf5-rtTA/TetO-cre/R26R-Confetti mice at 8 (**f**) and 20 (**g**) weeks post-induction in adulthood and immunostained for the luminal marker E-cadherin (blue). Arrow depicts an alveolar bud. Scale bars, 200 μm (whole-mounts) and 30 μm (insets, optical sections).

Elf5-labelled cells evident after a 20-week chase indicated depletion of the reservoir, consistent with the properties of progenitor rather than stem cells (Fig. 2g).

Elf5-expressing cells contributed to the generation of mature alveolar cells in pregnancy, with the emergence of sporadically labelled alveoli (Extended Data Fig. 4). The majority (~60%) of labelled alveoli contained progeny expressing 2–4 fluorescent proteins, implicating multiple progenitor cells in the generation of each alveolus. Given the high labelling efficiency observed with the *Elf5* model, the stochastic labelling suggests that many alveoli are generated from progenitors that lie upstream of *Elf5*-expressing cells. Interestingly, ducts but not residual alveoli in involuting glands were largely devoid of *Elf5*-labelled cells, indicating that these progenitors do not have a prominent role in remodelling (Extended Data Fig. 4). Overall, these data indicate that each round of alveologenesis is driven by a new pool of luminal progenitor cells.

Population dynamics of K5-expressing cells in puberty

We next addressed the nature of cells labelled in puberty using the *K5*-driven model and the Confetti reporter. The sea of colour evident within the cap and body layers of TEBs at 2 days after a pulse of doxycycline suggested the activation of multiple independent progenitors (Fig. 3a). Labelled luminal and myoepithelial cells in the ducts at 2-weeks post-induction were readily distinguished by whole-mount confocal imaging (Fig. 3b, Extended Data Fig. 3e). Although the epithelial trees remained multicoloured after an 8-week chase, discrete regions containing unicoloured epithelial cells of both lineages could be discerned (Fig. 3c, d). Evaluation of the number of fluorescent colours per duct indicated a shift in distribution between puberty and adulthood. The reduced number of multicoloured ducts in K5-rtTA/TetO-cre/R26R-Confetti glands after an 8-week chase contrasts with that observed for the *Elf5*-driven model (Extended Data Fig. 3g, h) and suggests that clonal expansion has occurred in the maturing mammary gland.

Quantification of the proportion of labelled cells within the basal and luminal lineages of K5-rtTA/TetO-cre/R26R-tdTomato glands revealed a marked expansion during puberty. The presence of tdTomato$^+$ luminal and basal cells at 2 days, 2 weeks and 8 weeks post-induction suggested that they might be derived from a common progenitor, given that *K5* only marks basal/myoepithelial cells (Fig. 3e and Extended Data Fig. 1d–h). The rapid *in vivo* labelling of both lineages further indicated that at least a subset of basal *K5*-expressing cells are actively dividing. Indeed, EdU (5-ethynyl-2′-deoxyuridine)-positive cells co-localized with endogenous K5-positive cells and *K5*-derived red fluorescent protein (RFP)/tdTomato$^+$ cells in the TEBs and ducts (Extended Data Fig. 5). Moreover, rare cells (2–3 per TEB) could be captured that co-expressed K5, keratin 14 (K14, also known as Krt14) and Elf5, indicating a transient state that may precede asymmetric division. Notably, the tdTomato$^+$ luminal population expanded considerably between the 2-day and 8-week chase (3.0 to 20.6%). Labelling of both the luminal and basal populations was also evident in another model driven by the human *K5* promoter (K5-creERT2/R26R-yellow fluorescent protein (YFP))[28], in which YFP$^+$ cells were found distributed between the two lineage populations (Extended Data Fig. 6a, b).

Bipotent stem cells exist in the adult gland

To track the fate of *K5*-expressing cells at a clonal level and determine their contribution to ductal homeostasis, K5-rtTA/TetO-cre/R26R-Confetti mice were pulsed in adulthood. Single- and multicoloured regions were evident along the mammary ducts at 8 weeks post-induction, with higher magnification revealing discrete clonal domains (Fig. 4a, b and Extended Data Fig. 7). Although most clonal patches contained both cell lineages, there were differences in the proportion of each among the clones. To address the fate of independent marked cells, triple-transgenic mice were analysed at 1 week post-induction. Small clonal regions (Extended Data Fig. 7d, e) could be subdivided into myoepithelial-only and 'mixed' clones (which either contained equi-numbers of luminal and myoepithelial cells or were enriched for luminal cells). Importantly, no isolated luminal cells were observed. The presence of both cell lineages within the clones provides unequivocal evidence for the targeting of bipotent basal cells in the adult mammary gland.

Enumeration of mixed and myoepithelial clones at 1 and 8 weeks post-induction revealed a preponderance of the mixed type and an increase in clone size over time (Fig. 4c, d and Extended Data Fig. 8). A shift in cellular composition was also noted for the mixed clones, with a substantial increase in the proportion of luminal-rich clones at 8 weeks. The abundance of single myoepithelial cells evident after a 1-week chase suggested that they had not yet divided or were differentiated (Fig. 4d). Scanning of ducts in 3D revealed that some K5$^+$ cells in the basal layer were tethered to luminal cells, compatible with recent asymmetric division of the basal cell (Fig. 4e and Extended Data Fig. 7). Further evidence for bipotent *K5*-marked cells came from quantitation using the tdTomato reporter allele. Concomitant with the basal population, tdTomato$^+$ cells in the luminal population

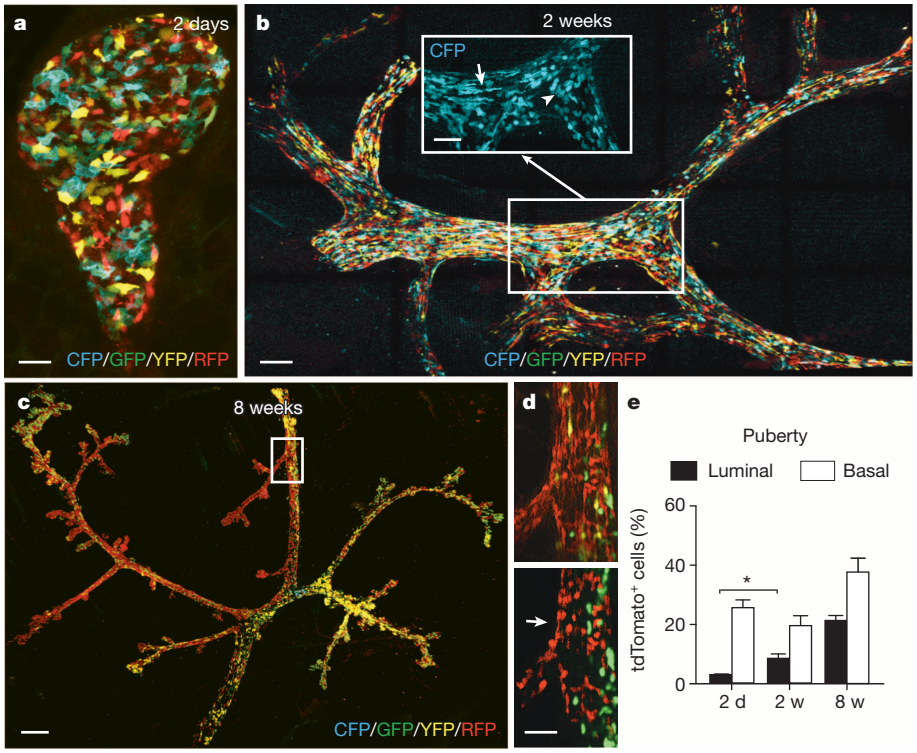


Figure 3 | K5-expressing cells labelled at the onset of puberty contribute to expansion of both epithelial lineages. **a**–**c**, Whole-mount 3D confocal images of a TEB or duct from K5-rtTA/TetO-cre/R26R-Confetti mice analysed at 2 days (**a**), 2 weeks (**b**) or 8 weeks (**c**) after doxycycline induction at the onset of puberty. Scale bars, 40 μm (**a**, TEB), 150 μm (**b**, **c**, ducts). The arrow and arrowhead (inset in **b**) depict myoepithelial and luminal cells, respectively. **d**, Enlargement of a clonal region showing optical sections of the myoepithelial (top panel) and luminal (bottom panel) layers. Arrow depicts luminal cells. Scale bars, 100 μm (optical sections). **e**, Bar chart showing the percentage of tdTomato$^+$ cells in the luminal and basal populations at the indicated times after pulsing in puberty. Data represent the mean ± s.e.m. for $n = 3$ mice. $^*P < 0.01$ (*t*-test).

markedly expanded from 1.7% at 2 days to 16.4% at 8 weeks post-induction in K5-rtTA/TetO-cre/R26R-tdTomato mice (Fig. 4f).

Unexpectedly, the epithelial tree appears to undergo active maintenance throughout adulthood. In the mammary glands of K5-rtTA/TetO-cre/R26R-Confetti mice subjected to a 20-week (Fig. 5a, b) or 52-week (data not shown) chase, extensive labelled domains were evident along many ducts. Quantification of labelled cells using the tdTomato reporter confirmed that both the luminal and basal populations were maintained (Fig. 4f). The profound segregation of colour in the whole-mount mammary glands after the extended chase period indicates clonal expansion. For example, the large ductal area appears to have been replenished by the progeny of two stem cells yielding large RFP$^+$ and YFP$^+$ clonal domains comprising myoepithelial and luminal cells (Fig. 5a, b). Thus, long-lived bipotent stem cells are capable of considerable

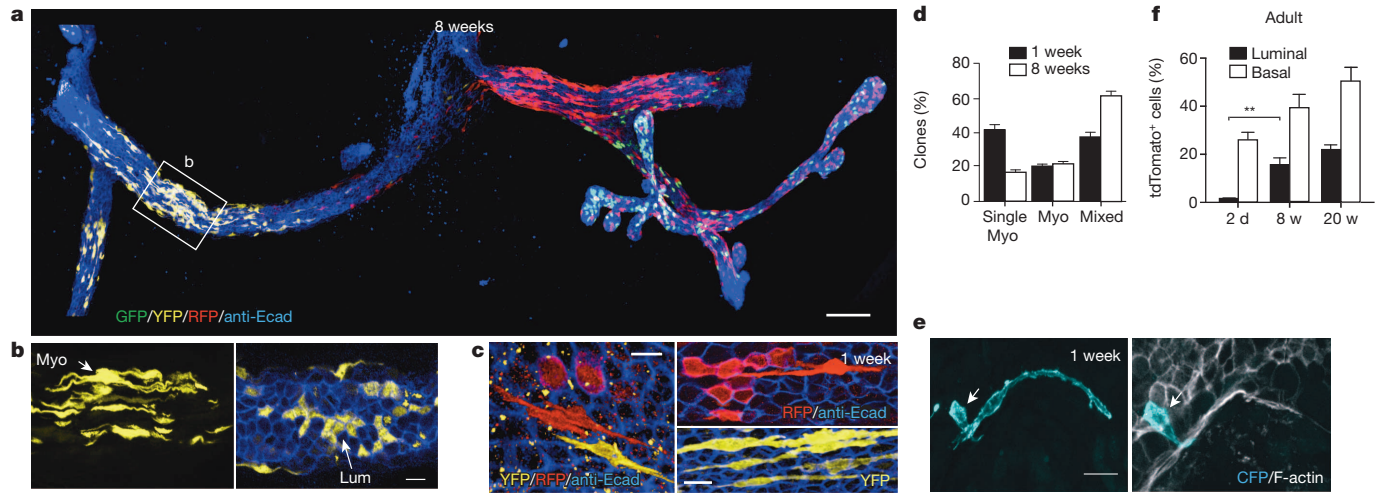


Figure 4 | K5-expressing bipotent stem cells ensure expansion of both lineages in the adult gland. **a**, Whole-mount 3D confocal image of ducts in K5-rtTA/TetO-cre/R26R-Confetti glands at 8 weeks post-doxycycline injection in adulthood and immunostained for E-cadherin (blue). **b**, Optical sections of the rectangular box in **a** showing luminal and myoepithelial layers of a clonal domain. Scale bars, 100 μm (**a**, whole-mount), 10 μm (**b**, optical sections). **c**, Whole-mount confocal images of clonal patches in K5-rtTA/TetO-cre/R26R-Confetti mammary glands at 1 week post-induction in adulthood, immunostained for E-cadherin (blue). Scale bars, 8 μm. **d**, Bar chart showing percentage of single myoepithelial cells (single Myo), myoepithelial (Myo) and mixed clones at 1 and 8 weeks post-induction: more than 1,000 (>200 mammary ducts) and 200 clones were counted at 1 week ($n = 5$ mice) and 8 weeks ($n = 3$ mice), respectively. **e**, Optical sections from a whole-mounted duct stained for F-actin (white) (Extended data Fig. 7) showing a myoepithelial cell (outer layer) attached to a luminal cell: left panel, 5-μm thick optical section; right panel, 0.43-μm section, indicating the luminal cell. Scale bar, 10 μm. **f**, Bar chart showing the percentage of tdTomato-expressing cells in the luminal and basal populations of mammary glands from K5-rtTA/TetO-cre/R26R-tdTomato mice at 2 days ($n = 3$), 8 ($n = 4$) and 20 weeks ($n = 3$), after induction in adulthood. In **d** and **f**, data represent the mean ± s.e.m. $^{**}P < 0.002$ for the luminal subset (*t*-test).

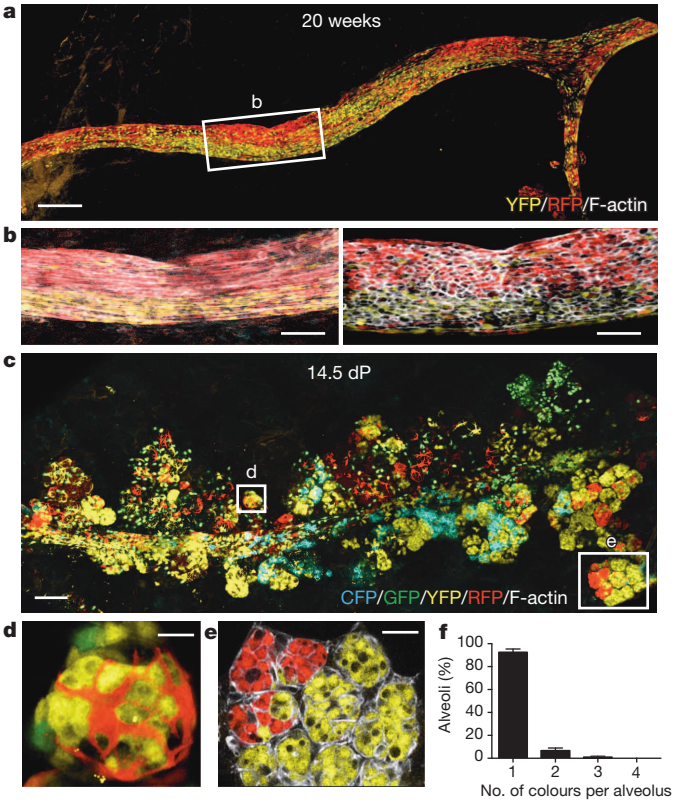


Figure 5 | K5 targets long-lived stem cells that contribute to maintenance and pregnancy. **a**, Whole-mount 3D confocal image of mammary ducts from K5-rtTA/TetO-cre/R26R-Confetti mice at 20 weeks after doxycycline injection in adulthood, stained for F-actin (white). **b**, Enlargement of the rectangle in **a** shows optical sections of the myoepithelial (left) and luminal layers (right). Scale bars, 100 μm (**a**, whole-mount) and 50 μm (**b**, optical sections). **c**–**e**, Whole-mount confocal image of alveoli in K5-rtTA/TetO-cre/R26R-Confetti glands at 14.5 days of pregnancy (14.5 dP), stained for F-actin (**c**), with enlargements shown in **d**, **e**: one alveolus comprising YFP^+ luminal and RFP^+/GFP^+ myoepithelial cells (**d**) and optical section of an alveolar unit comprising multicoloured alveoli (RFP or YFP) (**e**). Scale bars, 100 μm (**c**, whole-mount) and 13 μm (**d**) and 30 μm (**e**). **f**, Bar chart showing percentage of alveoli that exhibit 1, 2, 3 or 4 colours per alveolus for K5-rtTA/TetO-cre/R26R-Confetti mice at 14.5 days of pregnancy after pulsing in adulthood. Data represent mean ± s.e.m. ($n = 3$ mice) for 557 alveoli.

expansion and appear to coordinate ductal homeostasis. Moreover, the longevity of these cells makes them likely targets for the acquisition of mutations, which could result in breast cancer many years later.

K5-expressing cells also contribute to alveologenesis. Analysis of K5-rtTA/TetO-cre/R26R-Confetti glands at mid-pregnancy revealed labelled luminal cells within individual alveoli that were predominantly unicoloured, indicating their derivation from one stem cell (Fig. 5c–f). Moreover, alveolar and myoepithelial cells within a given alveolus often expressed different fluorescent proteins, indicating different origins (Fig. 5d). It seems probable that basal stem cells give rise to a pool of luminal progenitor cells, which then drive alveologenesis. Conversely, mature myoepithelial cells may migrate from the ducts to generate the discontinuous myoepithelial network around the alveolus. Thus, *K5*-labelled cells can contribute to all three mature cell types resident within the mammary gland. These data are consistent with the labelling of all epithelial cells by *K5*-Cre in the glands of Rosa26 mice[29] and cell-fate mapping studies suggesting that $Axin2^+$ cells contribute to alveologenesis[21].

Bipotent stem cells in *K14*- and *Lgr5*-driven models

To ascertain whether bipotent cells could be identified in other tracing models, we used the K14-creERT2 strain[30]. Given that oestrogen and its receptor have pivotal roles in mammary gland development, it was first necessary to determine the lowest possible dose of the anti-oestrogen tamoxifen to obviate any deleterious effects on ductal morphogenesis. A single, low dose of tamoxifen (1.5 mg), 5–10 times lower than that used in previous lineage-tracing studies[20], was selected. Notably, higher doses were shown to considerably impair mammary gland development, in contrast to doxycycline (Extended Data Fig. 9).

Analogous to findings obtained for *K5*-expressing cells, K14-creERT2/R26R-Confetti glands comprised multicoloured ducts and TEBs after administration of tamoxifen in puberty (Extended Data Fig. 10). Quantification of labelled cells in K14-creERT2 mice on the R26R-YFP background showed that YFP^+ cells were almost equally distributed between the luminal and basal populations (Extended Data Fig. 10). Pulsing of K14-creERT2/R26R-Confetti mice in adulthood revealed a multicoloured ductal tree with discrete clonal regions containing both cell lineages after an 8-week chase (Extended Data Fig. 10). Pertinently, mammary glands from mice subjected to two reproductive cycles were multicoloured, with discrete domains comprising both epithelial subtypes (Extended Data Fig. 11). Similar findings were made for the doxycycline-regulatable *K5* model after 2 weeks of involution. Collectively, these data indicate that bipotent stem cells have profound self-renewing capacity and that multiple stem cells are recruited for remodelling of the epithelial tree during involution.

Finally, we examined the contribution of the small basal subfraction of cells expressing Lgr5 to ductal maintenance and pregnancy by pulsing adult Lgr5-GFP-IRES-creERT2/R26R-tdTomato mice (ref. 31) (Extended Data Fig. 6). Clonal patches comprising both luminal and myoepithelial cells were occasionally visible along the ductal tree after an 8-week chase, consistent with FACS analysis showing labelling of both subsets. Furthermore, *Lgr5*-expressing cells contributed to the formation of alveoli, thus establishing their bipotential capacity. The low expression evident at this locus (and degree of labelling)[20,32] necessitates the use of 3D imaging for *in situ* detection of luminal cells. In agreement with previous findings[32], both *Lgr5*-GFP^+ and GFP^- cells exhibited repopulating potential upon transplantation at limiting dilution (1/110 versus 1/343, respectively), implying heterogeneity in the MaSC compartment. However, these data differ from another study in which lower numbers of $Lgr5^-$ cells were transplanted[33]. Thus, *Lgr5*-expressing cells can contribute to both cell lineages *in situ* and in regenerative assays, indicating that this gene can target bipotent precursor cells.

Discussion

The existence of bipotent stem cells has been the subject of intense debate in the mammary gland field given the apparent disparity between transplantation versus lineage-tracing assays[20,21]. It has been postulated that bipotent stem cells detected in the embryo no longer exist or function in the postnatal animal[20]. This study provides direct evidence for bipotent stem cells in the adult mammary gland where they have a fundamental role both in morphogenesis and homeostasis. High-resolution cell-fate mapping in four different transgenic models revealed clonal epithelial patches that comprised both myoepithelial and luminal cells and hence their derivation from a common basal precursor. Importantly, these cells do not merely serve as 'reserve' stem cells that are only activated upon damage or transplantation, but have an important physiological role in the postnatal gland. The full potential of such stem cells, however, may only be manifest *in vivo* at a specific window of development.

Bipotent stem cells are long lived and actively contribute to maintenance of ductal architecture. Extensive clonal domains populated by bipotent stem cells became evident upon long-term tracing and suggested that many ducts are maintained by the progeny of a few stem cells with extensive self-renewal capacity. The labelling patterns suggest a model whereby bipotent MaSCs initiate the replenishment process, which is then continued by unipotent progenitors generated from stem cells that have undergone asymmetric division. Although the luminal progenitors identified *in situ* appear to be long lived, their gradual depletion in the ageing gland and during involution indicates

that they are distinct from stem cells. The widespread distribution of *Elf5*-marked progenitors throughout the mammary tree during puberty suggests an important role in ductal morphogenesis. The similar array of multicoloured cells evident in the *K5*- and *K14*-driven models during puberty suggests that large numbers of progenitors derived from stem cells orchestrate ductal expansion.

Our observations differ from previous mammary lineage-tracing studies that used keratin gene promoters in different mouse strains and found long-lived basal-restricted cells throughout development[20]. Some inherent limitations of lineage tracing probably account for these differences. First, both transgenic and knock-in strains can be prone to insertion-site effects that affect the level and timing of gene expression, and ultimately determine whether *cre* expression is triggered in the stem or progenitor subset. A second consideration that is particularly relevant to the mammary gland is tamoxifen-associated toxicity. It is essential to use a low dose of this oestrogen receptor antagonist to prevent inactivation of the oestrogen receptor signalling axis in both luminal and stem cells[14]. Third, this technique is reliant on the use of highly specific promoters that accurately mirror expression of the endogenous gene in a given cell subtype. In cases in which the labelling efficiency is low, the fate of the majority of cells within the lineage and the fidelity of the promoter come into question. Finally, two-dimensional confocal analysis does not provide topological information. On the basis of the longevity of the progenitor described here, the unipotent cells tracked in previous studies may represent progenitor cells. Ultimately, a more precise picture of the mammary hierarchy will require cell-fate mapping studies together with the continued prospective isolation of refined subsets. The definition of these cellular subsets should provide a useful framework for identifying cells-of-origin and potential biomarkers in breast cancer.

METHODS SUMMARY

K5-rtTA-IRES-GFP and Elf5-rtTA-IRES-GFP transgenic mice were generated and crossed with TetO-cre and either R26R-Confetti[26] or R26R-tdTomato reporter strains to generate triple-transgenic mice. Cre-mediated excision was induced by doxycycline (2 mg) during puberty or adulthood. Similarly, reporter gene expression was induced in K14-creERT2 mice[30] (carrying R26R-Confetti or R26R-YFP), K5-creERT2/R26R-YFP[28] or Lgr5-creERT2/R26R-tdTomato mice by a single injection of tamoxifen (1.5 mg). Single-cell suspensions of mammary glands were prepared and analysed by FACS as described previously[8]. To prepare whole-mount mammary glands for 3D imaging, fat pads were briefly fixed in 4% paraformaldehyde and large sections of ductal network up to 3 mm in length were dissected. Native fluorescence was detected (Confetti mice) or immunohistochemical staining was performed on glands before dissection and imaging. For whole-mounted glands, tile scans of *Z*-stacks (around 200 μm thickness) were acquired and data sets analysed by Imaris with the 3D visualization module.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 August 2013; accepted 12 December 2013.
Published online 26 January 2014.

1. Hennighausen, L. & Robinson, G. W. Information networks in the mammary gland. *Nature Rev. Mol. Cell Biol.* **6,** 715–725 (2005).
2. Hoshino, K. & Gardner, W. U. Transplantability and life span of mammary gland during serial transplantation in mice. *Nature* **213,** 193–194 (1967).
3. Daniel, C. W., De Ome, K. B., Young, J. T., Blair, P. B. & Faulkin, L. J. Jr. The *in vivo* life span of normal and preneoplastic mouse mammary glands: a serial transplantation study. *Proc. Natl Acad. Sci. USA* **61,** 53–60 (1968).
4. Smith, G. H. & Medina, D. A morphologically distinct candidate for an epithelial stem cell in mouse mammary gland. *J. Cell Sci.* **90,** 173–183 (1988).
5. Asselin-Labat, M. L. *et al.* Gata-3 is an essential regulator of mammary-gland morphogenesis and luminal-cell differentiation. *Nature Cell Biol.* **9,** 201–209 (2007).
6. Eirew, P. *et al.* A method for quantifying normal human mammary epithelial stem cells with *in vivo* regenerative ability. *Nature Med.* **14,** 1384–1389 (2008).
7. Lim, E. *et al.* Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nature Med.* **15,** 907–913 (2009).
8. Shackleton, M. *et al.* Generation of a functional mammary gland from a single stem cell. *Nature* **439,** 84–88 (2006).
9. Shehata, M. *et al.* Phenotypic and functional characterization of the luminal cell hierarchy of the mammary gland. *Breast Cancer Res.* **14,** R134 (2012).
10. Sleeman, K. E., Kendrick, H., Ashworth, A., Isacke, C. M. & Smalley, M. J. CD24 staining of mouse mammary gland cells defines luminal epithelial, myoepithelial/basal and non-epithelial cells. *Breast Cancer Res.* **8,** R7 (2006).
11. Sleeman, K. E. *et al.* Dissociation of estrogen receptor expression and *in vivo* stem cell activity in the mammary gland. *J. Cell Biol.* **176,** 19–26 (2007).
12. Stingl, J. *et al.* Purification and unique properties of mammary epithelial stem cells. *Nature* **439,** 993–997 (2006).
13. Villadsen, R. *et al.* Evidence for a stem cell hierarchy in the adult human breast. *J. Cell Biol.* **177,** 87–101 (2007).
14. Asselin-Labat, M. L. *et al.* Control of mammary stem cell function by steroid hormone signalling. *Nature* **465,** 798–802 (2010).
15. Cicalese, A. *et al.* The tumor suppressor p53 regulates polarity of self-renewing divisions in mammary stem cells. *Cell* **138,** 1083–1095 (2009).
16. dos Santos, C. O. *et al.* Molecular hierarchy of mammary differentiation yields refined markers of mammary stem cells. *Proc. Natl Acad. Sci. USA* **110,** 7123–7130 (2013).
17. Smith, G. H. Label-retaining epithelial cells in mouse mammary gland divide asymmetrically and retain their template DNA strands. *Development* **132,** 681–687 (2005).
18. Joshi, P. A. *et al.* Progesterone induces adult mammary stem cell expansion. *Nature* **465,** 803–807 (2010).
19. Kretzschmar, K. & Watt, F. M. Lineage tracing. *Cell* **148,** 33–45 (2012).
20. Van Keymeulen, A. *et al.* Distinct stem cells contribute to mammary gland development and maintenance. *Nature* **479,** 189–193 (2011).
21. van Amerongen, R., Bowman, A. N. & Nusse, R. Developmental stage and time dictate the fate of Wnt/β-catenin-responsive stem cells in the mammary gland. *Cell Stem Cell* **11,** 387–400 (2012).
22. Lim, E. *et al.* Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. *Breast Cancer Res.* **12,** R21 (2010).
23. Oakes, S. R. *et al.* The Ets transcription factor Elf5 specifies mammary alveolar cell fate. *Genes Dev.* **22,** 581–586 (2008).
24. Grachtchouk, M. *et al.* Basal cell carcinomas in mice arise from hair follicle stem cells and multiple epithelial progenitor populations. *J. Clin. Invest.* **121,** 1768–1781 (2011).
25. Rinkevich, Y., Lindau, P., Ueno, H., Longaker, M. T. & Weissman, I. L. Germ-layer and lineage-restricted stem/progenitors regenerate the mouse digit tip. *Nature* **476,** 409–413 (2011).
26. Snippert, H. J. *et al.* Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143,** 134–144 (2010).
27. Mascré, G. *et al.* Distinct contribution of stem and progenitor cells to epidermal maintenance. *Nature* **489,** 257–262 (2012).
28. Rock, J. R. *et al.* Basal cells as stem cells of the mouse trachea and human airway epithelium. *Proc. Natl Acad. Sci. USA* **106,** 12771–12775 (2009).
29. Moumen, M. *et al.* The proto-oncogene Myc is essential for mammary stem cell function. *Stem Cells* **30,** 1246–1254 (2012).
30. Li, M. *et al.* Skin abnormalities generated by temporally controlled RXRα mutations in mouse epidermis. *Nature* **407,** 633–636 (2000).
31. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene *Lgr5*. *Nature* **449,** 1003–1007 (2007).
32. de Visser, K. E. *et al.* Developmental stage-specific contribution of LGR5$^+$ cells to basal and luminal epithelial lineages in the postnatal mammary gland. *J. Pathol.* **228,** 300–309 (2012).
33. Plaks, V. *et al.* Lgr5-expressing cells are sufficient and necessary for postnatal mammary gland organogenesis. *Cell Rep* **3,** 70–78 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are grateful to F. Jackling and K. Liu for genotyping, F. Vaillant for performing transplants, C. Nowell and K. Roger in the WEHI Imaging Facility for expert support, S. Firth at Monash MicroImaging, Leica and Zeiss for imaging support, D. Sieiro Mosti for help with quantification, J. Stanley for generation of transgenic strains, and the Animal, FACS and Histology facilities at WEHI. We also thank J. Adams for review of the manuscript, and H. Clevers, P. Chambon, M. Furtado, B. Hogan and M. Shen for the provision of mouse strains. This work was supported by the Australian National Health and Medical Research Council (NHMRC); the Victorian State Government through VCA funding of the Victorian Breast Cancer Research Consortium and Operational Infrastructure Support; the Australian Cancer Research Foundation; and the Qualtrough Family Bequest. A.C.R. and N.Y.F. were supported by a National Breast Cancer Foundation/Cure Cancer Australia Fellowship, G.J.L. by a NHMRC Research Fellowship and J.E.V. by an Australia Fellowship.

Author Contributions A.C.R. and N.Y.F. designed and performed all experiments, and carried out data analysis. J.E.V. and G.J.L. conceived the study and designed experiments. J.E.V., N.Y.F., A.C.R. and G.J.L. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.E.V. (visvader@wehi.edu.au).

ARTICLE

doi:10.1038/nature13039

Protein–guided RNA dynamics during early ribosome assembly

Hajin Kim[1,2]*†, Sanjaya C. Abeysirigunawarden[3]*, Ke Chen[4,5], Megan Mayerle[6]†, Kaushik Ragunathan[4]†, Zaida Luthey-Schulten[1,4,5], Taekjip Ha[1,2,4,5] & Sarah A. Woodson[3,6]

The assembly of 30S ribosomes requires the precise addition of 20 proteins to the 16S ribosomal RNA. How early binding proteins change the ribosomal RNA structure so that later proteins may join the complex is poorly understood. Here we use single-molecule fluorescence resonance energy transfer (FRET) to observe real-time encounters between *Escherichia coli* ribosomal protein S4 and the 16S 5′ domain RNA at an early stage of 30S assembly. Dynamic initial S4–RNA complexes pass through a stable non-native intermediate before converting to the native complex, showing that non-native structures can offer a low free-energy path to protein–RNA recognition. Three-colour FRET and molecular dynamics simulations reveal how S4 changes the frequency and direction of RNA helix motions, guiding a conformational switch that enforces the hierarchy of protein addition. These protein-guided dynamics offer an alternative explanation for induced fit in RNA–protein complexes.

The ribosome is a large cellular complex that synthesizes proteins. During assembly of the small (30S) subunit of the *E. coli* ribosome, 20 ribosomal proteins associate with the 16S ribosomal RNA (rRNA) in a defined hierarchy[1–3] that arises from protein-induced changes in the structure of the rRNA[4]. Despite progress in visualizing ribosome assembly intermediates[5], the physical basis for cooperative assembly is poorly understood because it depends on short-lived conformational states[6]. The simplest model is that early binding proteins capture the native structure of a helix junction[7–9], pre-organizing adjacent helices so that additional proteins can join the complex[10,11]. However, time-resolved footprinting of 30S assembly showed that some ribosomal proteins contact their rRNA binding site in stages[12], indicating that proteins do not capture the folded structure of the rRNA, but remodel it over time. Remodelling of early protein–RNA interactions has important implications for further stages of assembly.

To understand how proteins remodel the rRNA structure, we probed the motions between the rRNA and ribosomal protein S4 (Fig. 1) in real time using single-molecule FRET (smFRET)[13]. smFRET was previously used to observe the Mg^{2+}- or ribosomal protein S15-dependent conformational change of a three-helix rRNA element of the 30S ribosome[7]. Here, we use two- and three-colour FRET to determine the direction of helix motions as S4 binds a 542-nucleotide rRNA fragment.

S4 is one of the first proteins to bind the rRNA[14], and nucleates 30S ribosome assembly[15] by folding a five-way junction (5WJ) in the 16S 5′ domain (Fig. 1a). Footprinting and mutagenesis results showed that the S4–5′ domain complex recapitulates the native 30S protein–rRNA interactions[16,17] and is a substrate for further steps of assembly. S4 binding stabilizes both the folded 5WJ[17] and a conserved pseudoknot between helix (h) 18 and its internal loop (teal; Fig. 1a, b)[18,19] that is crucial for translation fidelity[20]. Conserved nucleotides in the h18 internal loop only fold correctly after S4 has bound[18,19].

RNA helix dynamics in S4–rRNA complexes

To observe internal motions in the S4–rRNA complex, we labelled S4 with a donor fluorophore, cyanine 3 (Cy3), via an engineered single cysteine (Methods). We also labelled the 5′ domain RNA by annealing a Cy5-conjugated oligonucleotide to a 3′ extension of 16S h3 helix (5′dom-h3). We labelled h3 because it docks under the h18 pseudoknot and contacts S4 in the mature 30S ribosome (Fig. 1b), yet was proposed to point away from h18 and S4 in an assembly intermediate[21]. Therefore, a label on h3 was likely to capture the dynamics of early assembly intermediates. Chemical footprinting and ensemble binding assays established that these modifications did not noticeably change the folding of the 5′ domain RNA or its affinity for S4 (Extended Data Figs 1–3), which is similar to that of the natural 16S rRNA[17].

Complexes of S4–Cy3 and 5′dom-h3–Cy5 were immobilized on a polymer-passivated quartz surface via biotin on the oligonucleotide extending from h3, and imaged by total internal reflection fluorescence microscopy. Single complexes over time displayed discrete transitions between two stable FRET states (Fig. 1c). Analysis of the dwell times showed that the low- and high-FRET states interconvert in 1–10 s in 20 mM Mg^{2+} (Extended Data Fig. 4). In 20 mM Mg^{2+}, S4 remained bound to the RNA and the complex stayed mostly in the high-FRET state (FRET efficiency $E \sim 0.7$) (Fig. 1c, d). Because native interactions bring the Cy5 acceptor on h3 close to the Cy3 donor on S4 (ref. 22), we inferred that the high-FRET state represents the docked conformation of h3 observed in the 30S ribosome, which we take to be the native state of this complex. At 4 mM Mg^{2+}, we observed frequent dissociation and re-binding of S4 (Fig. 1c) and greater occupancy of the low-FRET state ($E \sim 0.2$) (Fig. 1d). We assigned this low-FRET state to a 'flipped' assembly intermediate in which h3 has swung away from the body of the complex, in agreement with footprinting of the S4–rRNA complex[21].

[1]Department of Physics, Center for the Physics of Living Cells and Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. [2]Howard Hughes Medical Institute, Urbana, Illinois 61801, USA. [3]T. C. Jenkins Department of Biophysics, Johns Hopkins University, 3400 N. Charles Street, Baltimore, Maryland 21218, USA. [4]Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. [5]Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. [6]CMDB Program, Johns Hopkins University, 3400 N. Charles Street, Baltimore, Maryland 21218, USA. †Present addresses: School of Nano-Bioscience and Chemical Engineering, Ulsan National Institute of Science and Technology, Ulsan 689-798, Republic of Korea (H.K.); Department of Biochemistry and Biophysics, University of California at San Francisco, 600 16th Street, San Francisco, California 94143-2200, USA (M.M.); Department of Cell Biology, Harvard Medical School, 240 Longwood Avenue, LHRRB-517, Boston, Massachusetts 02115-5730, USA (K.R.).
*These authors contributed equally to this work.

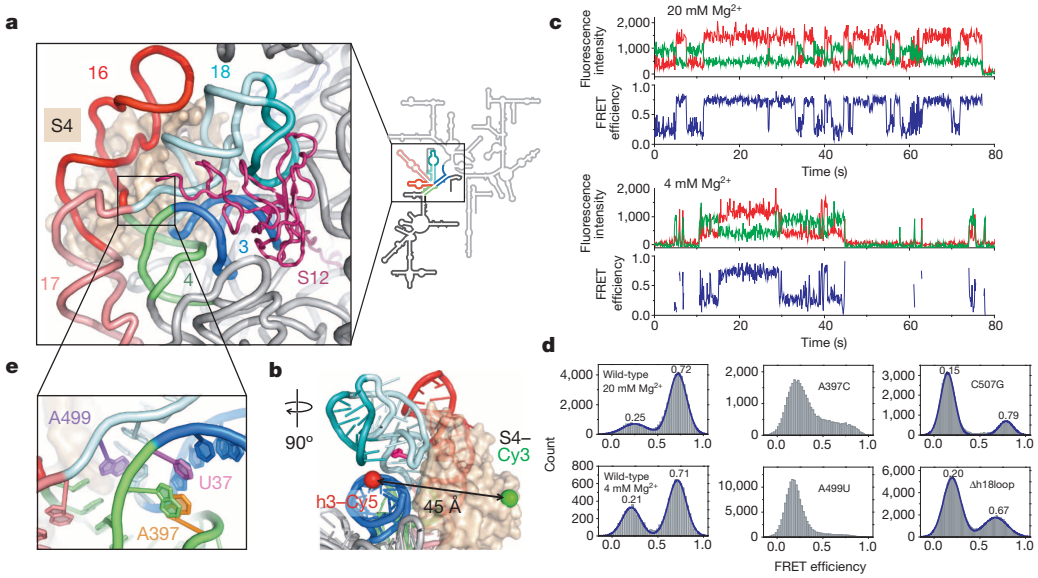


Figure 1 | Fluctuations during early ribosome assembly. **a**, Protein S4 (tan surface) bound to the 16S rRNA in the *E. coli* 30S ribosome (PDB accession 2I2P). 5WJ coloured in the ribbon and 16S schematic; rest of 5′ domain, black. S12 (dark red) binds the 50S interface side of 5WJ in the mature 30S ribosome. **b**, Labelling positions for FRET between S4–Cy3 (green sphere) and 5′dom-h3–Cy5 (red sphere). C507 in the h18 pseudoknot is in magenta. **c**, smFRET traces show two-state fluctuations in the distance between S4 and h3. Green, Cy3 fluorescence; red, Cy5 fluorescence; blue, FRET efficiency. The trace at 4 mM Mg^{2+} exhibits frequent S4 binding and dissociation. **d**, FRET histograms of wild-type complex in 20 and 4 mM Mg^{2+} and mutant RNAs in 20 mM Mg^{2+}. Δh18loop lacks the teal region in **a**. **e**, Stacking interactions by A397 and A499 stabilize the h3–h18 junction in the ribosome.

These structural assignments were validated by rRNA mutations predicted to destabilize the docked conformation of h3 (Fig. 1e). The mutation A397C removes a tertiary interaction between h3 and h4, whereas the mutation A499U disrupts adenine stacking that maintains a 90° angle between h3 and h18 (ref. 23). These mutations did not reduce S4 binding, but they prevented stable docking of h3 in 20 mM Mg^{2+}, as revealed by a higher population of low-FRET structures and a broader distribution of mid-to-high FRET values (Fig. 1d and Extended Data Fig. 4). In addition, a base mismatch in the h18 pseudoknot (C507G; Fig. 1b), or deletion of the h18 loop (Δh18loop; teal in Fig. 1b) reduced the high-FRET population, showing that interactions with the folded h18 loop stabilize the docked (native) state of h3 (Fig. 1d and Extended Data Fig. 4).

A dynamic encounter complex

We next asked how S4 initially binds the rRNA. Time-resolved footprinting experiments showed that S4 contacts the 16S rRNA in multiple stages spanning 10 ms to 100 s (ref. 12). These stretched binding dynamics may arise from conformational changes in the 16S rRNA[22] or co-folding of the S4 amino terminus, which interacts extensively with h16 and h18 in the ribosome yet is mostly unstructured in free S4 (ref. 24).

We measured the dynamics of S4 binding by flowing S4–Cy3 over immobilized 5′dom-h3–Cy5 RNA pre-folded in 20 mM Mg^{2+} (Fig. 2a). An abrupt increase in fluorescence defined the moment of initial contact with S4–Cy3 (Fig. 2b and Extended Data Fig. 5). We built a map of the FRET distribution over time by synchronizing single-molecule traces to the moment of binding (Fig. 2c). Although the high-FRET state was favoured at equilibrium in 20 mM Mg^{2+}, the FRET distribution at the moment of binding (histogram in Fig. 2c) showed that more than 80% of complexes passed through the low-FRET state before switching to the high-FRET state in ~1 s.

In 4 mM Mg^{2+}, closer to the physiological condition, ≥40% of complexes showed transient spikes of fluorescence, meaning that S4 dissociated before the complexes could progress to the native state (Fig. 2d and Extended Data Fig. 5). A synchronized map of successful binding trajectories showed that the FRET values at the initial encounter were broadly distributed and included a mid-FRET state (E ~0.6) that was rarely observed in 20 mM Mg^{2+} (Fig. 2e). Lowering [Mg^{2+}] to 2 mM shifted the initial FRET values further towards ~0.6 (Fig. 2f). Analysis of individual trajectories showed that these diverse initial complexes first transitioned to the low-FRET (flipped) state before reaching the high-FRET (docked) state (Fig. 2d and Extended Data Fig. 5). This sequential pathway was evident in the synchronized map that showed a transient build-up of low-FRET population in the first 1–2 s after initial association (Fig. 2e, f and Extended Data Fig. 6).

To better visualize initial S4 encounters, we made additional observations with 10-ms resolution. These experiments revealed that initial complexes fluctuate rapidly between different structures (Fig. 2g; see also Extended Data Fig. 7) before converging to the low-FRET state within ~0.2 s (Fig. 2h, orange rectangle). Although the S4–rRNA complex starts from a disordered ensemble, it does not directly jump to its native structure but progresses through the low-FRET intermediate. Thus, S4 interactions steer the system from a heterogeneously fluctuating encounter complex towards a slower dynamic equilibrium between two RNA conformations.

Kinetic model of S4–rRNA binding pathway

Our observations of individual binding events revealed a minimal kinetic pathway for S4–rRNA recognition that accounts for the stretched timeframe of RNP assembly (Fig. 3a). In our model, a rapidly fluctuating, heterogeneous S4–5′ domain encounter complex (EC) diffuses to a less-dynamic, low-FRET flipped intermediate (FI), in which h3 swings away from S4 and h18. Finally, h3 docks against S4 and underneath the h18 pseudoknot to form the high-FRET native complex (NC).

The lifetimes of these states in different [Mg^{2+}] provide additional information on the S4 binding mechanism (Fig. 3b–g). First, we estimated the binding rate constant (k_{on}) by observing the delay between the addition of S4–Cy3 and the appearance of a fluorescent signal on the rRNA. The maximum k_{on} ~5×10^7 M^{-1} s^{-1} agreed well with time-resolved footprinting[12], showing that initial binding occurs near the diffusion limit for macromolecules. S4 bound the 5′ domain RNA fastest in 10 mM Mg^{2+}, which in footprinting experiments[22] favours the flipped conformation of h3, but 16 times slower in 20 mM Mg^{2+}, which favours the docked conformation of h3 (Fig. 3b). This supports our observation that the non-native flipped conformation offers a low free-energy path for S4 binding, and contrasts with the usual assumption that proteins are more likely to recognize the natively folded RNA.

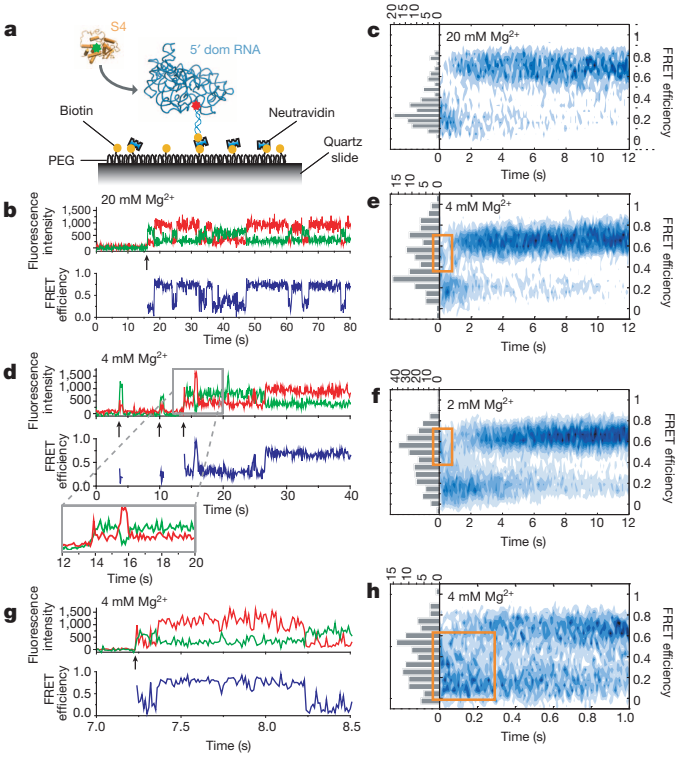


Figure 2 | Dynamics of S4 binding to the 5′ domain RNA. **a**, S4–Cy3 was added to immobilized 5′ domain RNA labelled with Cy5 at h3. PEG, polyethylene glycol. **b**, **d**, Binding traces in 20 and 4 mM Mg^{2+}, respectively. Arrows indicate when S4–Cy3 binds. **c**, **e**, **f**, Probability density maps of synchronized FRET dynamics. Histograms show FRET distribution at the moment of binding. Orange rectangles denote initial mid-FRET population. **g**, Binding trace at 4 mM Mg^{2+} acquired at 10-ms resolution. **h**, Synchronized FRET density map at 10-ms resolution. Orange rectangle denotes the broad initial FRET distribution converging to the low-FRET state.

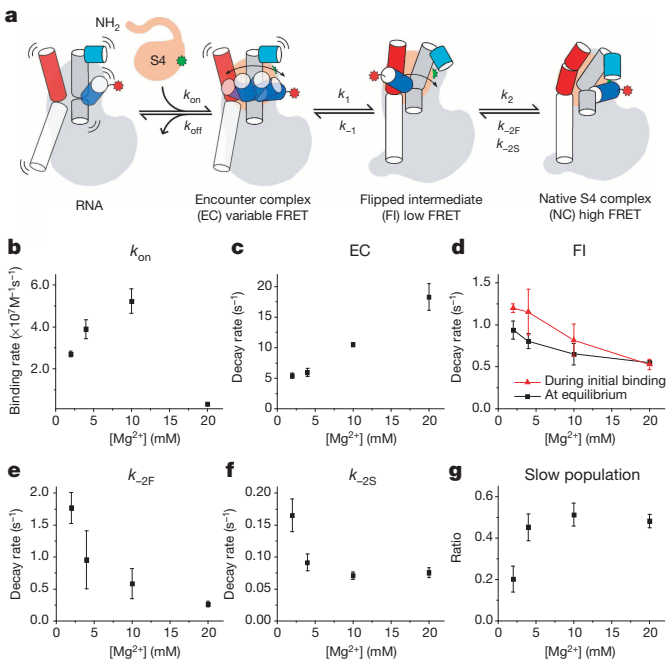


Figure 3 | Kinetic pathway for S4 binding. **a**, Model constructed from single-molecule data. **b**–**f**, $[Mg^{2+}]$-dependence of rate constants for initial binding (**b**), decay of the EC (**c**), FI (**d**) and NC (**e**, **f**) via fast (k_{-2F}, **e**) and slow (k_{-2S}, **f**) components. **g**, The fraction of the slow component, $A_S/(A_S+A_F)$, when fitting the NC decay as $A_S e^{-k_{-2S}t}+A_F e^{-k_{-2F}t}$. A_S and A_F are the amplitudes of each component; k_{-2S} and k_{-2F} are the decay rate constants. The decay rate of FI state was comparable between initial binding measurements (red) and equilibrium dynamics (black). Error bars, ± s.e.m. from triplicate measurements (see Methods for data statistics).

Second, Mg^{2+} ions aid transitions from EC to FI, based on faster decay of EC at higher $[Mg^{2+}]$ (Fig. 3c). Conversely, transitions between the FI and NC states slowed down with increasing $[Mg^{2+}]$ (Fig. 3d–f), implying that Mg^{2+} ions stabilize the flipped and docked complexes more than the transition state between them. Reorientation of h3 may require partial unstacking of the h3–h18 junction and cation release, as observed in DNA and RNA four-way junctions[25,26]. Finally, the longer lifetime of NC at higher $[Mg^{2+}]$ (Fig. 3f–g) explains why Mg^{2+} ions move the equilibrium towards NC[27]. In the cell, where free $[Mg^{2+}]$ is ~1 mM, other ribosomal proteins may help to stabilize NC[21].

S4 guides the RNA dynamics

We next investigated how S4 changes pre-existing motions in the free RNA. To compare the conformational dynamics of the rRNA before and after S4 binding, we introduced a second label in the rRNA at the tip of h16 (5′dom-h3h16). Previous smFRET measurements of a minimal 5WJ revealed the complex folding energy landscape of the RNA in the absence of S4 (ref. 28). We observed similarly heterogeneous dynamics between h3 and h16 in the larger 5′ domain RNA, with varied transition rates between low- and high-FRET states in different molecules (Fig. 4a and Extended Data Fig. 8). These changes in RNA structure are reversible as a subset of molecules switched their dynamics within a single trajectory (Extended Data Fig. 9), as observed in other RNAs[29–34].

To see how S4 modulated RNA dynamics, we performed three-colour smFRET[35] experiments between S4–Cy3, h16–Cy5 and h3–Cy7. By alternating Cy3 and Cy5 excitation, we measured all pairwise distances pseudo-simultaneously, yielding information on the direction as well as frequency of motion[35,36]. S4 binding suppressed relative motions of h16, as inferred from the lack of discernible fluctuations in the S4–h16 and h16–h3 FRET efficiencies (Fig. 4b). The FRET from S4 to h3 still fluctuated between high and low values, consistent with two-colour data showing that h3 continues to move after S4 binds (Fig. 1c).

We used the observed FRET efficiencies to construct a geometric model of how S4 modulates the motions of the RNA helices. Without S4, the distance between h3 and h16 fluctuates heterogeneously (black arrows in Fig. 4c) as expected for the loosely folded free RNA[17]. When S4 binds, persistent low FRET between h16 and S4 shows that the motion of h16 is restricted and agrees with the expected distance between these labels in the 30S ribosome (Fig. 4d). Conversely, steady low FRET between h3 and h16, combined with fluctuations in the S4–h3 distance, can be explained by movement of h3 in a plane perpendicular to h16, such that its distance to h16 does not change substantially (blue arrows in Fig. 4c). These dynamics are not likely due to motions in S4, because solution NMR showed that the Cy3-labelled domain of S4 is stably folded[37].

Molecular dynamics simulations

To gain further insight into how S4 modulates the dynamics of the RNA, we performed all-atom molecular dynamics simulations of a minimal RNA containing just the 5WJ (Fig. 4e). We pictorialized the global motions of the helices by tracing the movements of their tips (Methods). The helix motions in these and other simulations of the 5WJ in 0–30 mM Mg^{2+} (ref. 28) agreed well with the experimentally observed S4–h3–h16 distances. Without S4, h16 explored two distinct regions of space whereas h3 swept out a wide cone (Fig. 4e, top panel), consistent with the fluctuations in FRET between h3 and h16 in the free rRNA (Fig. 4a). S4 fixed h16, allowing it to sample only a small region around its native structure (Fig. 4e, bottom panel), consistent with the stable S4–h16 FRET signal (Fig. 4b). More interestingly, S4 constrained h3 to an in-plane bending motion towards and away from S4, maintaining a nearly constant distance between h3 and h16.

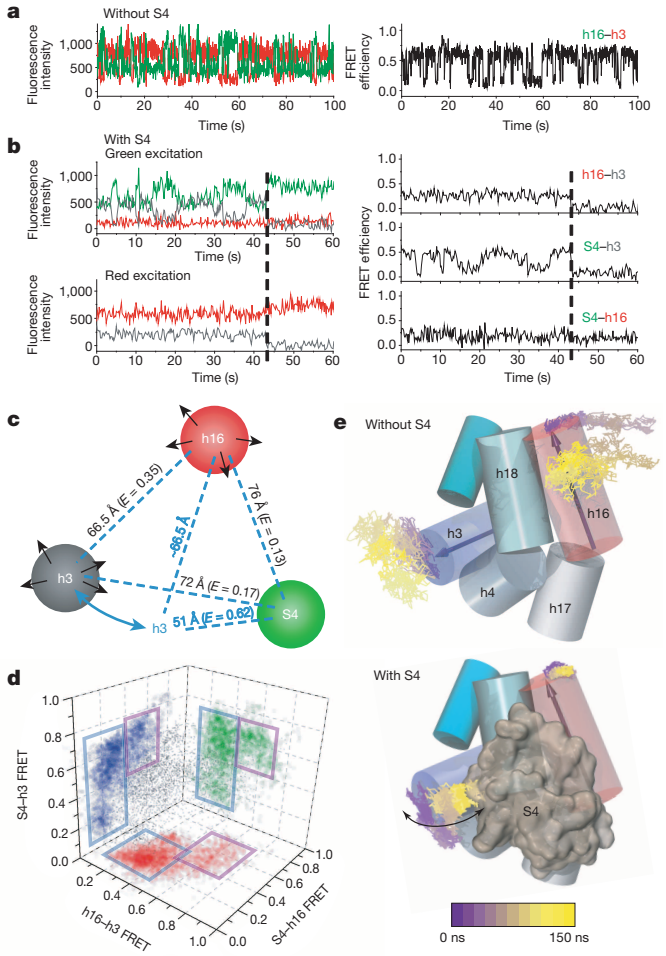


Figure 4 | Modulation of the rRNA dynamics by S4 binding. **a**, Dynamics of the free rRNA in 20 mM Mg^{2+}, on the basis of fluctuation in the h16–Cy3 to h3–Cy5 distance. **b**, Three-colour FRET shows that the h16–h3 dynamics are suppressed by bound S4, whereas S4–h3 dynamics are observed. **c**, Schematic model based on the observed distances and dynamics. Colours of spheres represent the labels for three-colour measurements (S4–Cy3, h16–Cy5, h3–Cy7). Isotropic motions of h16 and h3 in the free RNA are denoted by black arrows. Restricted motion of h3 when S4 binds is denoted by blue arrows. **d**, Distances between S4, h16 and h3 from 35 stable complexes (grey dots) were projected on each face to show the correlation between each pair of distances. Most molecules behave as in **c** (blue rectangles); a minor misfolded population shows higher FRET for h16–h3 and S4–h16 and stable high FRET for S4–h3 (violet rectangles)[28]. **e**, RNA dynamics from 150-ns all-atom molecular dynamics simulation with and without S4. Thin lines trace movement (from violet to yellow) of h16 (red cylinder) and h3 (blue cylinder) during the simulation.

Separate hybrid MD-Gō simulations of non-equilibrium binding also reproduced the switch from flipped to docked conformations of h3 and the fluctuating encounter conformation at mid-FRET (Extended Data Fig. 10a, b). In most MD-Gō trajectories, initial contact with the carboxy-terminal domain of S4 quickly reduced motions in the 5WJ. This allowed h16 to interact with the disordered N-terminal domain of S4 in the second stage of binding (Extended Data Fig. 10c). Therefore, both experiments and simulations showed that protein S4 alters the frequency and the range of motions in the rRNA: not only are fluctuations in the 5WJ damped as expected, immobilizing h16, but large-scale motions of h3 are confined to a specific direction.

Discussion

Protein-induced remodelling of RNA structures is widespread and assists the hierarchical assembly of RNA–protein complexes[4,11,38,39]. Unexpectedly, the low free-energy path for S4–rRNA recognition passes from a mobile encounter complex through a non-native intermediate before reaching the natively folded complex. S4 not only slows rRNA fluctuations; it induces anisotropic motions between the intermediate and the native complex. These protein-induced changes in the rRNA dynamics are reminiscent of the classical model of substrate-induced fit in enzymes[40]. Induced fit has been proposed as a universal feature in RNA–protein recognition, as both protein and RNA structures often change upon binding[41]. Our results suggest an alternative picture of induced fit, in which S4 changes the ensemble of thermally accessible RNA conformations by modulating the RNA dynamics. This differs from binding of small ligands to RNA helix junctions, which trap conformations accessed by bending motions of the unliganded RNA[42]. The ability of S4 to change the RNA dynamics after it binds is probably due to its larger interface with the 16S rRNA, and to the plasticity of its N-terminal domain. Many ribosomal proteins contain flexible segments that change structure upon RNA binding[43]. Transient protein–RNA interactions may facilitate the search for a high-affinity configuration[44–46].

We anticipate that the observed S4-guided dynamics of the rRNA contribute to subsequent steps of 30S assembly. First, the conserved h18 loop interacts with transfer RNA in the mature ribosome, and must fold correctly for normal protein synthesis. h18 folds correctly only after S4 binds the 16S rRNA in a temperature-dependent step[18]. Our experiments show that folded h18 stabilizes the docked form of h3 (Extended Data Fig. 4). Native interactions between h18 and h3 enable folding of h1 (Fig. 1a) and the central pseudoknot at the end of 30S assembly[47,48]. S4 mutations that inhibit remodelling of h18 and h3 docking impair 30S assembly in *E. coli*[49].

Second, h3 dynamics affect the recruitment of other proteins to the 16S rRNA. Protein S16, which adds to 30S complexes only when S4 is present, stabilizes the docked conformation of h3 (ref. 21). Later in 30S assembly, protein S12 binds the docked h3 on the side opposite S4 (Fig. 1a). The restricted in-plane motions of h3 between the flipped and docked conformations may reject S12 until the h18 loop has folded correctly and S16 has bound, ensuring the proper hierarchy of protein addition. Thus, protein-guided RNA dynamics creates additional checkpoints for molecular interactions, thereby improving the fidelity of rRNA folding and ribosome assembly.

METHODS SUMMARY

The 5′ domain of *E. coli* 16S rRNA (nucleotides (nt) 21–556) was labelled by hybridizing dye-labelled oligonucleotides to an extension of h3 (5′dom-h3) or expansion of the h16 loop (5′dom-h3h16) (Extended Data Table 1). *E. coli* protein S4(C32S, S189C) labelled with Cy3 was incubated with the RNA in 80 mM K-HEPES, pH 7.6, 300 mM KCl, plus the desired [$MgCl_2$] for 5 min at 37 °C, or flowed into slide chambers containing immobilized RNA for real-time binding experiments. Complexes were immobilized on quartz slides for imaging with a total internal reflection fluorescence microscope (100 ms per frame except where stated otherwise). Two-colour FRET efficiencies were defined with leakage correction as $E_{FRET} = (I_A - 0.06 \times I_D)/(I_D + I_A)$, where I_D and I_A are the apparent fluorescence intensities of the donor and acceptor, respectively. Three-colour FRET efficiencies were corrected for the leakage, direct excitation of acceptor dyes by lasers, and the wavelength-dependent detection efficiency. See full Methods for data statistics. Molecular dynamics simulations (150 ns) using NAMD2 and CHARMM27 force fields started from 5WJ RNA and S4 models based on the *E. coli* ribosome structure (PDB 2I2P), neutralized with Mg^{2+} and Na^{+} ions, and solvated with a periodic boundary condition. Hybrid MD-Gō simulations of 100 S4 binding trajectories started from an extended RNA structure and included a Lennard-Jones Gō potential to drive protein–RNA association[28].

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 September 2013; accepted 20 January 2014.
Published online 12 February 2014.

1. Held, W. A., Ballou, B., Mizushima, S. & Nomura, M. Assembly mapping of 30S ribosomal proteins from *Escherichia coli*. Further studies. *J. Biol. Chem.* **249,** 3103–3111 (1974).

2. Nierhaus, K. H. & Dohme, F. Total reconstitution of functionally active 50S ribosomal subunits from *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **71,** 4713–4717 (1974).
3. Traub, P. & Nomura, M. Structure and function of *Escherichia coli* ribosomes. VI. Mechanism of assembly of 30S ribosomes studied *in vitro*. *J. Mol. Biol.* **40,** 391–413 (1969).
4. Stern, S., Powers, T., Changchien, L. M. & Noller, H. F. RNA-protein interactions in 30S ribosomal subunits: folding and function of 16S rRNA. *Science* **244,** 783–790 (1989).
5. Shajani, Z., Sykes, M. T. & Williamson, J. R. Assembly of bacterial ribosomes. *Annu. Rev. Biochem.* **80,** 501–526 (2011).
6. Woodson, S. A. RNA folding pathways and the self-assembly of ribosomes. *Acc. Chem. Res.* **44,** 1312–1319 (2011).
7. Ha, T. *et al.* Ligand-induced conformational changes observed in single RNA molecules. *Proc. Natl Acad. Sci. USA* **96,** 9077–9082 (1999).
8. Weeks, K. M. & Cech, T. R. Assembly of a ribonucleoprotein catalyst by tertiary structure capture. *Science* **271,** 345–348 (1996).
9. Caprara, M. G., Mohr, G. & Lambowitz, A. M. A tyrosyl-tRNA synthetase protein induces tertiary folding of the group I intron catalytic core. *J. Mol. Biol.* **257,** 512–531 (1996).
10. Agalarov, S. C., Sridhar Prasad, G., Funke, P. M., Stout, C. D. & Williamson, J. R. Structure of the S15,S6,S18-rRNA complex: assembly of the 30S ribosome central domain. *Science* **288,** 107–112 (2000).
11. Kuglstatter, A., Oubridge, C. & Nagai, K. Induced structural changes of 7SL RNA during the assembly of human signal recognition particle. *Nature Struct. Biol.* **9,** 740–744 (2002).
12. Adilakshmi, T., Bellur, D. L. & Woodson, S. A. Concurrent nucleation of 16S folding and induced fit in 30S ribosome assembly. *Nature* **455,** 1268–1272 (2008).
13. Ha, T. *et al.* Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc. Natl Acad. Sci. USA* **93,** 6264–6268 (1996).
14. Talkington, M. W., Siuzdak, G. & Williamson, J. R. An assembly landscape for the 30S ribosomal subunit. *Nature* **438,** 628–632 (2005).
15. Nowotny, V. & Nierhaus, K. H. Assembly of the 30S subunit from *Escherichia coli* ribosomes occurs via two assembly domains which are initiated by S4 and S7. *Biochemistry* **27,** 7051–7055 (1988).
16. Stern, S., Wilson, R. C. & Noller, H. F. Localization of the binding site for protein S4 on 16S ribosomal RNA by chemical and enzymatic probing and primer extension. *J. Mol. Biol.* **192,** 101–110 (1986).
17. Bellur, D. L. & Woodson, S. A. A minimized rRNA-binding site for ribosomal protein S4 and its implications for 30S assembly. *Nucleic Acids Res.* **37,** 1886–1896 (2009).
18. Powers, T. & Noller, H. F. A temperature-dependent conformational rearrangement in the ribosomal protein S4.16S rRNA complex. *J. Biol. Chem.* **270,** 1238–1242 (1995).
19. Mayerle, M., Bellur, D. L. & Woodson, S. A. Slow formation of stable complexes during coincubation of minimal rRNA and ribosomal protein S4. *J. Mol. Biol.* **412,** 453–465 (2011).
20. Powers, T. & Noller, H. F. A functional pseudoknot in 16S ribosomal RNA. *EMBO J.* **10,** 2203–2214 (1991).
21. Ramaswamy, P. & Woodson, S. A. S16 throws a conformational switch during assembly of 30S 5′ domain. *Nature Struct. Mol. Biol.* **16,** 438–445 (2009).
22. Ramaswamy, P. & Woodson, S. A. Global stabilization of rRNA structure by ribosomal proteins S4, S17, and S20. *J. Mol. Biol.* **392,** 666–677 (2009).
23. Grabow, W. W., Zhuang, Z., Swank, Z. N., Shea, J. E. & Jaeger, L. The right angle (RA) motif: a prevalent ribosomal RNA structural pattern found in group I introns. *J. Mol. Biol.* **424,** 54–67 (2012).
24. Sayers, E. W., Gerstner, R. B., Draper, D. E. & Torchia, D. A. Structural preordering in the N-terminal region of ribosomal protein S4 revealed by heteronuclear NMR spectroscopy. *Biochemistry* **39,** 13602–13613 (2000).
25. Hohng, S. *et al.* Conformational flexibility of four-way junctions in RNA. *J. Mol. Biol.* **336,** 69–79 (2004).
26. McKinney, S. A., Declais, A. C., Lilley, D. M. & Ha, T. Structural dynamics of individual Holliday junctions. *Nature Struct. Biol.* **10,** 93–97 (2003).
27. Gerstner, R. B., Pak, Y. & Draper, D. E. Recognition of 16S rRNA by ribosomal protein S4 from Bacillus stearothermophilus. *Biochemistry* **40,** 7165–7173 (2001).
28. Chen, K. *et al.* Assembly of the five-way junction in the ribosomal small subunit using hybrid MD-Go simulations. *J. Phys. Chem. B* **116,** 6819–6831 (2012).
29. Zhuang, X. *et al.* Correlating structural dynamics and function in single ribozyme molecules. *Science* **296,** 1473–1476 (2002).
30. Tan, E. *et al.* A four-way junction accelerates hairpin ribozyme folding via a discrete intermediate. *Proc. Natl Acad. Sci. USA* **100,** 9308–9313 (2003).
31. Xie, Z., Srividya, N., Sosnick, T. R., Pan, T. & Scherer, N. F. Single-molecule studies highlight conformational heterogeneity in the early folding steps of a large ribozyme. *Proc. Natl Acad. Sci. USA* **101,** 534–539 (2004).
32. Ditzler, M. A., Rueda, D., Mo, J., Hakansson, K. & Walter, N. G. A rugged free energy landscape separates multiple functional RNA folds throughout denaturation. *Nucleic Acids Res.* **36,** 7088–7099 (2008).
33. Solomatin, S. V., Greenfeld, M., Chu, S. & Herschlag, D. Multiple native states reveal persistent ruggedness of an RNA folding landscape. *Nature* **463,** 681–684 (2010).
34. Haller, A., Altman, R. B., Souliere, M. F., Blanchard, S. C. & Micura, R. Folding and ligand recognition of the TPP riboswitch aptamer at single-molecule resolution. *Proc. Natl Acad. Sci. USA* **110,** 4188–4193 (2013).
35. Hohng, S., Joo, C. & Ha, T. Single-molecule three-color FRET. *Biophys. J.* **87,** 1328–1337 (2004).
36. Munro, J. B., Altman, R. B., Tung, C. S., Sanbonmatsu, K. Y. & Blanchard, S. C. A fast dynamic mode of the EF-G-bound ribosome. *EMBO J.* **29,** 770–781 (2010).
37. Markus, M. A., Gerstner, R. B., Draper, D. E. & Torchia, D. A. The solution structure of ribosomal protein S4 Δ41 reveals two subdomains and a positively charged surface that may interact with RNA. *EMBO J.* **17,** 4559–4571 (1998).
38. Rose, M. A. & Weeks, K. M. Visualizing induced fit in early assembly of the human signal recognition particle. *Nature Struct. Biol.* **8,** 515–520 (2001).
39. Stone, M. D. *et al.* Stepwise protein-mediated RNA folding directs assembly of telomerase ribonucleoprotein. *Nature* **446,** 458–461 (2007).
40. Koshland, D. E. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl Acad. Sci. USA* **44,** 98–104 (1958).
41. Williamson, J. R. Induced fit in RNA–protein recognition. *Nature Struct. Biol.* **7,** 834–837 (2000).
42. Zhang, Q., Stelzer, A. C., Fisher, C. K. & Al-Hashimi, H. M. Visualizing spatially correlated dynamics that directs RNA conformational transitions. *Nature* **450,** 1263–1267 (2007).
43. Klein, D. J., Moore, P. B. & Steitz, T. A. The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *J. Mol. Biol.* **340,** 141–177 (2004).
44. Bokinsky, G. *et al.* Two distinct binding modes of a protein cofactor with its target RNA. *J. Mol. Biol.* **361,** 771–784 (2006).
45. Rau, M., Stump, W. T. & Hall, K. B. Intrinsic flexibility of snRNA hairpin loops facilitates protein binding. *RNA* **18,** 1984–1995 (2012).
46. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chem. Biol.* **5,** 789–796 (2009).
47. Holmes, K. L. & Culver, G. M. Mapping structural differences between 30S ribosomal subunit assembly intermediates. *Nature Struct. Mol. Biol.* **11,** 179–186 (2004).
48. Clatterbuck Soper, S. F., Dator, R. P., Limbach, P. A. & Woodson, S. A. *In vivo* X-ray footprinting of pre-30S ribosomes reveals chaperone-dependent remodeling of late assembly intermediates. *Mol. Cell* **52,** 506–516 (2013).
49. Mayerle, M. & Woodson, S. A. Specific contacts between protein S4 and ribosomal RNA are required at multiple stages of ribosome assembly. *RNA* **19,** 574–585 (2013).

Acknowledgements This work was supported by grants from the National Institutes of Health (R01 GM60819 to S.A.W.; R01 GM65367 to T.H.) and from the National Science Foundation (NSF) (PHY0822613 to T.H. and MCB12-44570 to Z. L.-S.). Supercomputer computing time was provided by NSF XSEDE (TG-MCA03S027). T.H. is an investigator with the Howard Hughes Medical Institute.

Author Contributions H.K., S.C.A., Z.L.-S., T.H. and S.A.W. designed the research. H.K., S.C.A., K.R. and M.M. conducted experiments, S.C.A. and M.M. provided samples, K.C. performed molecular dynamics simulations, H.K. and K.R. analysed the data and H.K. and S.A.W. wrote the paper with input from other authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.H. (tjha@illinois.edu) or S.A.W. (swoodson@jhu.edu).

LETTER

doi:10.1038/nature12997

Asymmetries in core–collapse supernovae from maps of radioactive ^{44}Ti in Cassiopeia A

B. W. Grefenstette[1], F. A. Harrison[1], S. E. Boggs[2], S. P. Reynolds[3], C. L. Fryer[4], K. K. Madsen[1], D. R. Wik[5], A. Zoglauer[2], C. I. Ellinger[6], D. M. Alexander[7], H. An[8], D. Barret[9,10], F. E. Christensen[11], W. W. Craig[2,12], K. Forster[1], P. Giommi[13], C. J. Hailey[14], A. Hornstrup[11], V. M. Kaspi[8], T. Kitaguchi[15], J. E. Koglin[16], P. H. Mao[1], H. Miyasaka[1], K. Mori[14], M. Perri[13,17], M. J. Pivovaroff[12], S. Puccetti[13,17], V. Rana[1], D. Stern[18], N. J. Westergaard[11] & W. W. Zhang[5]

Asymmetry is required by most numerical simulations of stellar core-collapse explosions, but the form it takes differs significantly among models. The spatial distribution of radioactive ^{44}Ti, synthesized in an exploding star near the boundary between material falling back onto the collapsing core and that ejected into the surrounding medium[1], directly probes the explosion asymmetries. Cassiopeia A is a young[2], nearby[3], core-collapse[4] remnant from which ^{44}Ti emission has previously been detected[5–8] but not imaged. Asymmetries in the explosion have been indirectly inferred from a high ratio of observed ^{44}Ti emission to estimated ^{56}Ni emission[9], from optical light echoes[10], and from jet-like features seen in the X-ray[11] and optical[12] ejecta. Here we report spatial maps and spectral properties of the ^{44}Ti in Cassiopeia A. This may explain the unexpected lack of correlation between the ^{44}Ti and iron X-ray emission, the latter being visible only in shock-heated material. The observed spatial distribution rules out symmetric explosions even with a high level of convective mixing, as well as highly asymmetric bipolar explosions resulting from a fast-rotating progenitor. Instead, these observations provide strong evidence for the development of low-mode convective instabilities in core-collapse supernovae.

Titanium-44 is produced in Si burning in the innermost regions of the material ejected in core-collapse supernovae, in the same processes that produce Fe and ^{56}Ni (ref. 13). The decay of radioactive ^{44}Ti (in the decay chain $^{44}\mathrm{Ti} \rightarrow {}^{44}\mathrm{Sc} \rightarrow {}^{44}\mathrm{Ca}$) results in three emission lines of roughly equal intensity at 67.86, 78.36 and 1,157 keV. Previous detections of the 1,157-keV line by the COMPTEL[5] instrument on the Compton Gamma Ray Observatory and of the 67.86- and 78.36-keV lines by the satellite X-ray telescopes Beppo-SAX[6], RXTE[7] and INTEGRAL[8] were of relatively low statistical significance individually, but when combined[8] they indicate a flux in each of the 67.86 and 78.36 keV lines of $(2.3 \pm 0.3) \times 10^{-5}$ photons (ph) $\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. For an explosion date of AD 1671[2], a distance of 3.4 kpc (ref. 3) and a half-life of 60 yr (ref. 14), this translates into a synthesized ^{44}Ti mass of $1.6^{+0.6}_{-0.3} \times 10^{-4} M_{\odot}$, where $M_{\odot}$ is the solar mass. Because of the limited spectral and spatial resolution, previous observations are not able to constrain the line centroid or spatial distribution within the remnant, although the non-detection of the 1,157-keV line by INTEGRAL/SPI has been used to place a lower limit of 500 $\mathrm{km\,s}^{-1}$ on the line width.

The space-based Nuclear Spectroscopic Telescope Array (NuSTAR) high-energy X-ray telescope[15], which operates in the band from 3 to 79 keV, observed Cas A, the remnant of a type IIb supernova[4], for multiple epochs between August 2012 and June 2013 with a total exposure of 1.2 Ms (Extended Data Table 1). The spectrum (Fig. 1) shows two clear, resolved emission lines with centroids redshifted by ~0.5 keV relative to the rest-frame ^{44}Ti decays of 67.86 and 78.36 keV. The telescope optics response cuts off at 78.39 keV (owing to the Pt K edge in the reflective coatings), which may affect the measured line centroid, width and flux of the 78.36-keV line, and so we focus on the 67.86-keV line for quantitative analysis. All errors are given at 90% confidence unless otherwise stated. We measure a line flux of $1.51 \pm 0.31 \times 10^{-5}$ ph $\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, implying a ^{44}Ti yield of $(1.25 \pm 0.3) \times 10^{-4} M_{\odot}$. This confirms previous spatially integrated ^{44}Ti yield measurements with a high statistical significance (Methods). The ^{44}Ti line is redshifted by 0.47 ± 0.21 keV, corresponding to a bulk line-of-sight Doppler velocity of 1,100–3,000 $\mathrm{km\,s}^{-1}$. The line is also broadened with a Gaussian half-width at half-maximum of 0.86 ± 0.26 keV. Assuming a uniformly expanding sphere, the corresponding velocity for the fastest material is $5{,}350 \pm 1{,}610$ $\mathrm{km\,s}^{-1}$.

The spatial distribution of emission in the 65–70-keV band (Fig. 2 and Extended Data Fig. 1) shows that the ^{44}Ti is clumpy and is slightly extended along the 'jet' axis seen in in the X-ray Si/Mg emission[11] and fast-moving optical knots[12]. There are also knots (that is, compact regions of emission) clearly evident off the jet axis. There is no evident alignment of the emission opposite to the direction of motion of the compact central object (CCO) as might be expected if the CCO kick involves an instability at the accretion shock[16].

We find that at least 80% (Extended Data Fig. 2) of the observed ^{44}Ti emission is contained within the reverse-shock radius as projected on the plane of the sky. Assuming a ~5,000 $\mathrm{km\,s}^{-1}$ expansion velocity from above and an age of 340 yr, the fastest-moving, outermost material with the highest line-of-sight velocity is 1.8 ± 0.5 pc from the centre of the explosion, which is consistent with the 1.6-pc radius estimated for the reverse shock[17]. This rules out the possibility that the ^{44}Ti is elongated along the line of sight and exterior to the reverse shock and is only observed in the interior of the remnant due to projection effects. We conclude that a majority of the ^{44}Ti is in the unshocked interior.

A striking feature of the NuSTAR ^{44}Ti spatial distribution is the lack of correlation with the Fe K-shell emission measured by the Chandra X-ray observatory (Fig. 3). In a supernova explosion, incomplete Si burning produces ejecta enriched with a range of elements including Si and Fe, whereas 'pure' Fe ejecta result either from complete Si burning or from the α-particle-rich freeze-out process that also produces ^{44}Ti. Although the fraction of Fe in such pure ejecta is difficult to constrain observationally[18], most models predict that a significant fraction of the

[1]Cahill Center for Astrophysics, 1216 East California Boulevard, California Institute of Technology, Pasadena, California 91125, USA. [2]Space Sciences Laboratory, University of California, Berkeley, California 94720, USA. [3]Physics Department, North Carolina State University, Raleigh, North Carolina 27695, USA. [4]CCS-2, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA. [5]NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA. [6]Department of Physics, University of Texas at Arlington, Arlington, Texas 76019, USA. [7]Department of Physics, Durham University, Durham DH1 3LE, UK. [8]Department of Physics, McGill University, Rutherford Physics Building, Montreal, Quebec H3A 2T8, Canada. [9]Université de Toulouse, UPS-OMP, IRAP, 9 Avenue du Colonel Roche, BP 44346, 31028 Toulouse Cedex 4, France. [10]CNRS, Institut de Recherche en Astrophysique et Planétologie, 9 Avenue colonel Roche, BP 44346, F-31028 Toulouse Cedex 4, France. [11]DTU Space, National Space Institute, Technical University of Denmark, Elektrovej 327, DK-2800 Lyngby, Denmark. [12]Lawrence Livermore National Laboratory, Livermore, California 94550, USA. [13]Agenzia Spaziale Italiana (ASI) Science Data Center, Via del Politecnico snc, I-00133 Roma, Italy. [14]Columbia Astrophysics Laboratory, Columbia University, New York, New York 10027, USA. [15]RIKEN, Nishina Center, 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan. [16]Kavli Institute for Particle Astrophysics and Cosmology, SLAC National Accelerator Laboratory, Menlo Park, California 94025, USA. [17]INAF – Osservatorio Astronomico di Roma, via di Frascati 33, I-00040 Monteporzio, Italy. [18]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA.

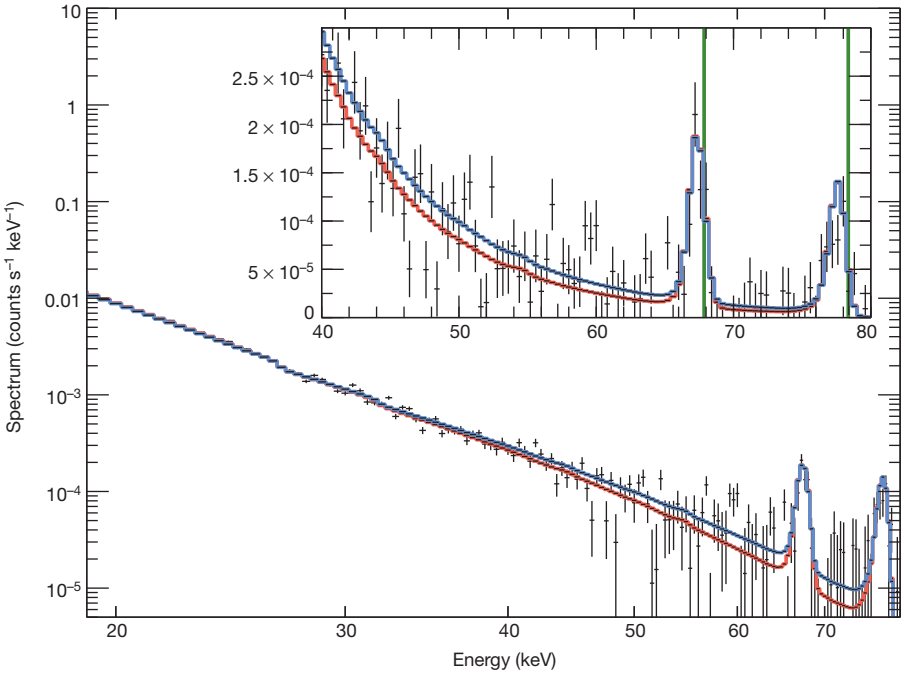


Figure 1 | The broadband hard-X-ray spectrum of Cas A. Data from both telescopes over all epochs are combined and shown as black data points with 1σ error bars. The spectra are shown combined and rebinned for plotting purposes only. Also shown are the best-fit continuum models for a power law (blue) and a model that describes electron cooling due to synchrotron losses (red). The continuum fits were obtained using the 10–60-keV data and extrapolated to 79 keV with the best-fit values for the continuum models provided in Extended Data Table 2, although the choice of continuum model does not significantly affect the measurement of the lines (Methods). When the continuum is extrapolated to 79 keV, clearly visible line features (Extended Data Fig. 5) appear near the ^{44}Ti line energies. Inset: zoomed region around the ^{44}Ti lines showing the data and the two models on a linear scale. The vertical green lines are the rest-frame energies of the ^{44}Ti lines (67.86 and 78.36 keV). A significant shift of ~0.5 keV to lower energy is evident for both lines, indicating a bulk line-of-sight velocity away from the observer. Details of the data analysis, including a discussion of the NuSTAR background features (Extended Data Fig. 4), are given in Methods. Extended Data Table 3 lists the parameters of the best-fit Gaussian models of these features with the error estimates described in Methods.

Fe is produced in close physical proximity to the ^{44}Ti. Some correlation would therefore be expected. The simplest explanation for the lack of correlation is that much of the Fe-rich ejecta have not yet been penetrated by the reverse shock and therefore do not radiate in the X-ray band. Whereas X-rays from ^{44}Ti decay are produced by a nuclear transition and directly trace the distribution of synthesized material, the Fe X-ray emission results from an atomic transition and traces the (mathematical) product of the Fe density and the density of shock-heated electrons; without the hot electrons, the Fe will not be visible in the X-rays. A possible explanation of our observations is that the bulk of the Fe ejecta in Cas A have not yet been shock-heated, further constraining models[18–20] of the remnant as well as the total amount of Fe. An alternative explanation is that most of the Fe is already shocked and visible, and that some mechanism decouples the production of ^{44}Ti and Fe and produces the observed uncorrelated spatial map.

Unshocked or cool, dense material (material that either was never heated or has already cooled after being shock-heated) might still be visible in the optical or infrared spectral band. The Spitzer space telescope observes line emission from interior ejecta primarily in [Si II] but it seems that there is not a significant amount of Fe present in these regions[21]. However, if unshocked ejecta are of sufficiently low density or have the wrong ionization states, then they will be invisible in the infrared and optical. Low-density Fe-rich regions may in fact exist interior to the reverse-shock radius as a result of inflation of the emitting material by radioactivity (the 'nickel bubble' effect[22]).

The concentration of Fe-rich ejecta inferred from maps in X-ray atomic transitions is well outside the region where it is synthesized, and not in the centre of the remnant interior to the reverse shock. This observation has been used to suggest the operation of a strong instability similar to that proposed for SN 1993J[23]. The presence of a significant fraction of the ^{44}Ti interior to the reverse shock and the implied presence of interior 'invisible' iron requires this conclusion be revisited.

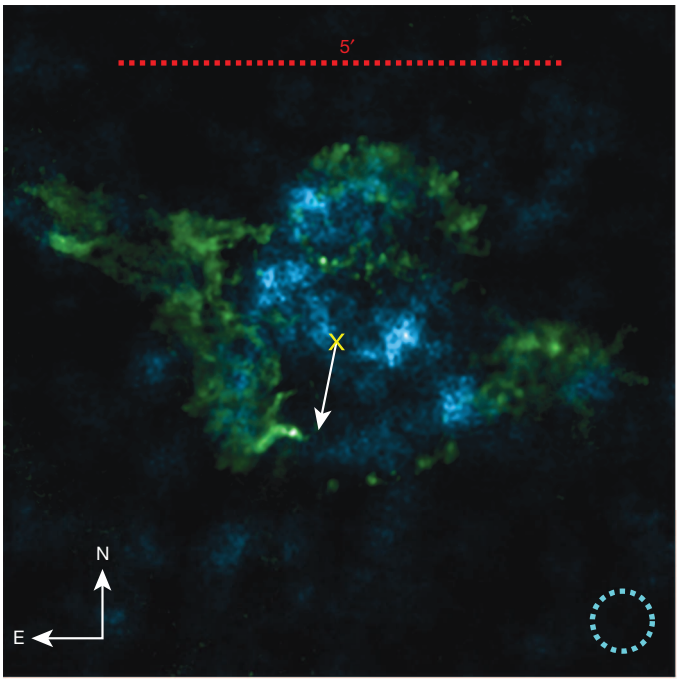


Figure 2 | A comparison of the spatial distribution of the ^{44}Ti with the known jet structure in Cas A. The image is oriented in standard astronomical coordinates as shown by the compass in the lower left and spans just over 5′ on a side. The ^{44}Ti observed by NuSTAR is shown in blue, where the data have been smoothed using a top-hat function with a radius shown in the lower right (dashed circle). The ^{44}Ti is clearly resolved into distinct knots and is non-uniformly distributed and almost entirely contained within the central 100″ (Methods and Extended Data Fig. 2). Shown for context in green is the Chandra ratio image of the Si/Mg band (data courtesy of NASA/CXC; Si/Mg ratio image courtesy of J. Vink), which highlights the jet–counterjet structure, the centre of the expansion of the explosion[2] (yellow cross) and the direction of motion of the compact object (white arrow). In contrast to the bipolar feature seen in the spatial distribution of Si ejecta, which argues for fast rotation or a jet-like explosion, the distribution of ^{44}Ti is much less elongated and contains knots of emission away from the jet axis. A reason for this may be that the Si originates in the outer stellar layers and is probably highly influenced by asymmetries in the circumstellar medium, unlike the ^{44}Ti, which is produced in the innermost layers near the collapsing core.

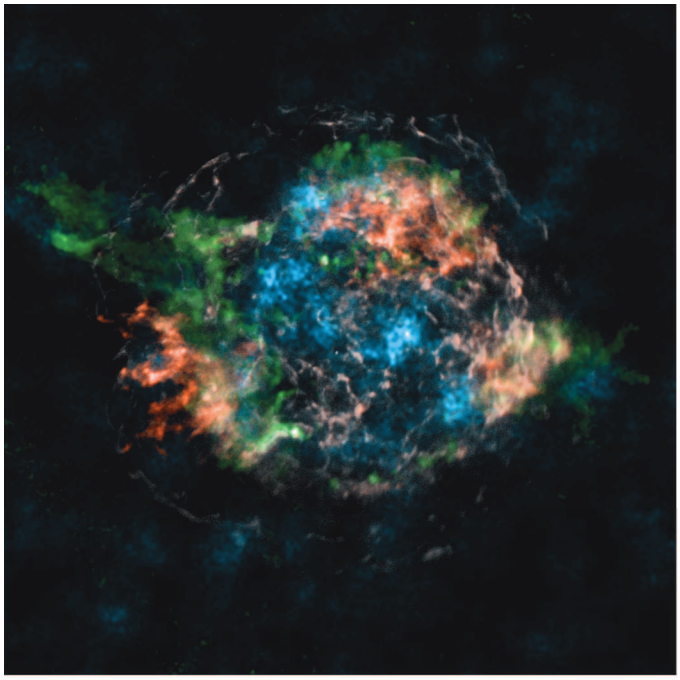


Figure 3 | A comparison of the spatial distribution of ^{44}Ti with known Fe K-shell emission in Cas A. We reproduce the spatial distributions shown in Fig. 2 and add the 4–6-keV continuum emission (white) and the spatial distribution of X-ray-bright Fe (red) seen by Chandra (Fe distribution courtesy of U. Hwang). We find that the ^{44}Ti does not follow the distribution of Fe K-shell X-ray emission, suggesting either that a significant amount of Fe remains unshocked and therefore does not radiate in the X-ray, or that the Fe/Ti ratio in the ejecta deviates from the expectation of standard nucleosynthesis models.

The measured ^{44}Ti line widths and distribution can directly constrain mixing in the supernova engine. As evidenced by SN 1987A, mixing due to Rayleigh–Taylor instabilities occurring between the explosion's forward and reverse shocks (distinct from the remnant's forward and reverse shocks) may be important in some types of supernova explosion[24]. Because ^{44}Ti is a good spatial tracer of ^{56}Ni in all established supernova models, we can compare the measured velocity width to that predicted for ^{56}Ni by simulations. We find that the $\sim$5,000 km s^{-1} maximum velocity and the level of Doppler line broadening compares well with type IIb models including mixing[25] and excludes models without the growth of large instabilities.

The evidence for asymmetries in the supernova explosion mechanism has grown steadily over the past several decades. Asymmetries are implied by a number of observations[26]: the extensive mixing implied in nearby supernovae (for example SN 1987A), the high space velocities of neutron stars and the polarization of supernova emission. Although different external processes could separately explain each of these observations, it is generally assumed that the asymmetries arise in the explosion mechanism. A number of mechanisms have been proposed within the framework where the supernova engine is convectively enhanced[27]: asymmetric collapse, asymmetries caused by rotation and asymmetries caused by low-mode convection. Of these, rotation and low-mode convection have received the most attention. Rotation tends to produce bipolar explosions along the rotation axis where the ejecta velocities are two to four times greater along this axis[28] than in the rest of the ejecta. Low-mode convection, including the standing accretion shock instability, will produce a bipolar explosion in fast-rotating stars, but is likely to produce higher-order modes in slowly rotating systems[29].

To improve our understanding of the nature of the observed ^{44}Ti non-uniformity, we compare the observations with three-dimensional models of normal core-collapse supernovae using a progenitor designed to produce the high ^{44}Ti/^{56}Ni ratio needed to match the estimated yields in the Cas A remnant. We simulate two explosions that represent the extremes of explosion asymmetry: a spherically symmetric explosion, and an explosion representing a fast-rotating progenitor with artificially induced bipolar asymmetry where the explosion velocity in a 30° half-angle cone near the rotation axis is increased by a factor of four relative to the rest of the ejecta. The simulated ^{44}Ti maps (Extended Data Fig. 3) indicate that the level of observed non-uniformity in Cas A is far greater than what can be produced by the spherically symmetric explosion, and that the bipolar explosion (where the bulk of the fast ^{44}Ti remains within 30° of the rotation axis) cannot reproduce the observed off-axis ^{44}Ti knots. This argues against fast-rotating progenitors as well as jet-like explosions, which are even more collimated than the bipolar explosions. The supernova is better described by an intermediate case, where the observed non-uniformity in the ^{44}Ti is the result of a multimodal explosion such as those predicted in both low-mode Rayleigh–Taylor models[29] and models including the standing accretion shock instability[30]. The Cas A remnant provides the first strong evidence that this low-mode convection must occur.

METHODS SUMMARY

A full description of the methods, including data analysis, background modelling, error estimates, and supernovae simulations can be found in Methods.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 August; accepted 13 December 2013.

1. Magkotsios, G. *et al.* Trends in ^{44}Ti and ^{56}Ni from core-collapse supernovae. *Astrophys. J. Suppl. Ser.* **191,** 66–95 (2010).
2. Thorstensen, J. R., Fesen, R. A. & van den Bergh, S. The expansion center and dynamical age of the galactic supernova remnant Cassiopeia A. *Astrophys. J.* **122,** 297–307 (2001).
3. Reed, J. E., Hester, J. J., Fabian, A. C. & Winkler, P. F. The three-dimensional structure of the Cassiopeia A supernova remnant. I. The spherical shell. *Astrophys. J.* **440,** 706–721 (1995).
4. Krause, O. *et al.* The Cassiopeia A supernova was of type IIb. *Science* **320,** 1195–1197 (2008).
5. Iyudin, A. F. *et al.* COMPTEL observations of Ti-44 gamma-ray line emission from Cas A. *Astron. Astrophys.* **284,** L1–L4 (1994).
6. Vink, J. *et al.* Detection of the 67.9 and 78.4 keV lines associated with the radioactive decay of ^{44}Ti in Cassiopeia A. *Astrophys. J.* **560,** L79–L82 (2001).
7. Rothschild, R. E. & Lingenfelter, R. E. Limits to the Cassiopeia A ^{44}Ti line flux and constraints on the ejecta energy and the compact source. *Astrophys. J.* **582,** 257–261 (2003).
8. Renaud, M. *et al.* The signature of ^{44}Ti in Cassiopeia A revealed by IBIS/ISGRI on INTEGRAL. *Astrophys. J.* **647,** L41–L44 (2006).
9. Nagataki, S., Hashimoto, M.-a., Sato, K., Yamada, S. & Mochizuki, Y. S. The high ratio of ^{44}Ti/^{56}Ni in Cassiopeia A and the axisymmetric collapse-driven supernova explosion. *Astrophys. J.* **492,** L45–L48 (1998).
10. Rest, A. *et al.* Direct confirmation of the asymmetry of the Cas A supernova with light echoes. *Astrophys. J.* **732,** 3 (2011).
11. Hwang, U. *et al.* A million second *Chandra* view of Cassiopeia A. *Astrophys. J.* **615,** L117–L120 (2004).
12. Fesen, R. A. *et al.* The expansion asymmetry and age of the Cassiopeia A supernova remnant. *Astrophys. J.* **645,** 283–292 (2006).
13. Woosley, S. E. & Weaver, T. A. The evolution and explosion of massive stars. II. Explosive hydrodynamics and nucleosynthesis. *Astrophys. J. Suppl. Ser.* **101,** 181–235 (1995).
14. Ahmad, I. *et al.* Improved measurement of the ^{44}Ti half-life from a 14-year long study. *Phys. Rev. C* **74,** 065803 (2006).
15. Harrison, F. A. *et al.* The Nuclear Spectroscopic Telescope ARray (NuSTAR) high-energy X-ray mission. *Astrophys. J.* **770,** 103 (2013).
16. Wongwathanarat, A., Janka, H.-T. & Müller, E. Hydrodynamical neutron star kicks in three dimensions. *Astrophys. J.* **725,** L106–L110 (2010).
17. Gotthelf, E. V. *et al.* Chandra detection of the forward and reverse shocks in Cassiopeia A. *Astrophys. J.* **552,** L39–L43 (2001).
18. Hwang, U. & Laming, J. M. A. Chandra X-ray survey of ejecta in the Cassiopeia A supernova remnant. *Astrophys. J.* **746,** 130 (2012).
19. Chevalier, R. A. & Oishi, J. Cassiopeia A and its clumpy presupernova wind. *Astrophys. J.* **593,** L23–L26 (2003).
20. Hwang, U. & Laming, J. M. Where was the iron synthesized in Cassiopeia A? *Astrophys. J.* **597,** 362–373 (2003).
21. Isensee, K. *et al.* The three-dimensional structure of interior ejecta in Cassiopeia A at high spectral resolution. *Astrophys. J.* **725,** 2059–2070 (2010).
22. Li, H., McCray, R. & Sunyaev, R. A. Iron, cobalt, and nickel in SN 1987A. *Astrophys. J.* **419,** 824–836 (1993).
23. Shigeyama, T. *et al.* Theoretical light curves of Type IIb supernova 1993J. *Astrophys. J.* **420,** 341–347 (1994).

24. Hachisu, I., Matsuda, T., Nomoto, K. i. & Shigeyama, T. Nonlinear growth of Rayleigh-Taylor instabilities and mixing in SN 1987A. *Astrophys. J.* **358,** L57–L61 (1990).
25. Nomoto, K. i., Iwamoto, K. & Suzuki, T. The evolution and explosion of massive binary stars and Type Ib-Ic-IIb-IIL supernovae. *Phys. Rep.* **256,** 173–191 (1995).
26. Hungerford, A. L., Fryer, C. L. & Warren, M. S. Gamma-ray lines from asymmetric supernovae. *Astrophys. J.* **594,** 390–403 (2003).
27. Janka, H.-T. Explosion mechanisms of core-collapse supernovae. *Annu. Rev. Nucl. Part. Sci.* **62,** 407–451 (2012).
28. Fryer, C. L. & Heger, A. Core-collapse simulations of rotating stars. *Astrophys. J.* **541,** 1033–1050 (2000).
29. Fryer, C. L. & Young, P. A. Late-time convection in the collapse of a 23 $M_{\odot}$ star. *Astrophys. J.* **659,** 1438–1448 (2007).
30. Blondin, J. M., Mezzacappa, A. & DeMarino, C. Stability of standing accretion shocks, with an eye toward core-collapse supernovae. *Astrophys. J.* **584,** 971–980 (2003).

Acknowledgements This work was supported by NASA under grant no. NNG08FD60C, and made use of data from the Nuclear Spectroscopic Telescope Array (NuSTAR) mission, a project led by Caltech, managed by the Jet Propulsion Laboratory and funded by NASA. We thank the NuSTAR operations, software and calibration teams for support with execution and analysis of these observations.

Author Contributions B.W.G.: reduction and modelling of the NuSTAR Cas A observations, interpretation, manuscript preparation. F.A.H.: NuSTAR principal investigator, observation planning, interpretation of results and manuscript preparation. S.E.B.: interpretation, manuscript review. S.P.R.: interpretation, manuscript preparation and review. C.L.F.: interpretation of results, manuscript review. K.K.M.: observation planning, data analysis, manuscript review. D.R.W.: background modelling, data analysis, manuscript review. A.Z.: background modelling, manuscript review. C.I.E.: supernova simulations, manuscript review. H.A.: image deconvolution, manuscript review. T.K.: detector modelling, data analysis, manuscript review. H.M., V.R., P.H.M.: detector production, response modelling, manuscript review. M.J.P.: optics calibration, manuscript review. S.P., M.P.: analysis software, calibration, manuscript review. K.F.: observation planning. F.E.C.: optics production and calibration, manuscript review. W.W.C.: optics and instrument production and response, observation planning, manuscript review. C.J.H.: optics production and response, interpretation, manuscript review. J.E.K.: optics production and response, manuscript review. N.J.W.: manuscript review, calibration. W.W.Z.: optics production and response, manuscript review. D.M.A., D.B., P.G., A.H., V.M.K., D.S.: science planning, manuscript review.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.W.G. (bwgref@srl.caltech.edu) or F.A.H. (fiona@srl.caltech.edu).

LETTER

doi:10.1038/nature13008

Fuel gain exceeding unity in an inertially confined fusion implosion

O. A. Hurricane[1], D. A. Callahan[1], D. T. Casey[1], P. M. Celliers[1], C. Cerjan[1], E. L. Dewald[1], T. R. Dittrich[1], T. Döppner[1], D. E. Hinkel[1], L. F. Berzak Hopkins[1], J. L. Kline[2], S. Le Pape[1], T. Ma[1], A. G. MacPhee[1], J. L. Milovich[1], A. Pak[1], H.-S. Park[1], P. K. Patel[1], B. A. Remington[1], J. D. Salmonson[1], P. T. Springer[1] & R. Tommasini[1]

Ignition is needed to make fusion energy a viable alternative energy source, but has yet to be achieved[1]. A key step on the way to ignition is to have the energy generated through fusion reactions in an inertially confined fusion plasma exceed the amount of energy deposited into the deuterium–tritium fusion fuel and hotspot during the implosion process, resulting in a fuel gain greater than unity. Here we report the achievement of fusion fuel gains exceeding unity on the US National Ignition Facility using a 'high-foot' implosion method[2,3], which is a manipulation of the laser pulse shape in a way that reduces instability in the implosion. These experiments show an order-of-magnitude improvement in yield performance over past deuterium–tritium implosion experiments. We also see a significant contribution to the yield from α-particle self-heating and evidence for the 'bootstrapping' required to accelerate the deuterium–tritium fusion burn to eventually 'run away' and ignite.

At the National Ignition Facility (NIF), 192 lasers deliver up to 1.9 MJ of light into a gold hohlraum, a cylindrically shaped radiation cavity (Fig. 1), that converts the energy into a nearly Planckian X-ray bath. A fraction of the X-rays are absorbed by a capsule generating ~100 Mbar of pressure in the ablator (the outer shell of the capsule). This ablation pressure, delivered as a series of weak shocks, accelerates the capsule inwards. Against the inside of the ablator is the deuterium–tritium (D–T) fuel shell, which is initially in a cryogenic ice state. When the implosion achieves peak velocity, the fuel has a kinetic energy that is a fraction of the X-ray energy absorbed by the capsule. As the fuel stagnates (stops moving), abruptly arrested by the high pressures forming at the centre of the implosion, the D–T forms a hotspot from the fuel's inner surface and PdV work (pressure times volume change) is done on the hotspot. The hotspot initiates the fusion reactions, producing neutrons and α-particles as the hotspot ion temperature climbs to many kiloelectronvolts. At sufficient hotspot areal density, $(\rho r)_{\rm hs} > 0.3\,{\rm g\,cm^{-2}}$, and ion temperature, $T_{\rm ion} > 4\,{\rm keV}$ (where Boltzmann's constant has been suppressed), the hotspot will 'ignite' as α-particles redeposit their energy locally. If the fuel areal density, $(\rho r)_{\rm fuel} > 1\,{\rm g\,cm^{-2}}$ the burn will propagate (heat adjacent cold fuel, causing it also to fuse and burn) and a run-away self-heating process releases energy many times greater than that absorbed by the capsule.

Mix of the ablator and D–T can degrade the ability of an inertially confined fusion implosion to compress the D–T fuel and can also cause undesirable cooling because high-atomic-number (high-Z) materials in the D–T hotspot will rapidly radiate away energy in the form of bremsstrahlung emission, the power of emission scaling in proportion to Z^2. Among many motivations, the high-foot implosion[2,3] was developed in the wake of the National Ignition Campaign[4,5] (NIC) primarily to address the possibility that ablation-front-driven instability[6,7] was responsible for part of the observed degraded yield performance[1] and the ablator–fuel mix inferred from X-ray emissions in combination with primary neutron yield[8,9].

The high-foot implosion is designed to reduce ablation-front-driven instability growth and thereby inhibit ablator plastic (carbon–hydrogen and silicon dopants) from mixing into and contaminating the D–T hotspot. The laser pulse shape is designed to obtain a relatively high hohlraum radiation temperature ($T_{\rm rad} \approx 90$–$100\,{\rm eV}$) during the 'foot' of the pulse (Fig. 1) and launches three shocks. In contrast, the NIC implosion pulse shape drives a lower radiation temperature ($T_{\rm rad} \approx 60\,{\rm eV}$) in the foot (hence 'low-foot') for longer and launches four shocks. The essential stability benefits of the high-foot scheme can be understood from examining an expression for the linear growth rate of the ablation-driven Rayleigh–Taylor instability[10]

$$\gamma_{\rm A-RTI} = \alpha_2({\rm Fr},\, v)\sqrt{\frac{kg}{1+kL_\rho}} - \beta_2({\rm Fr},\, v)kv_{\rm a} \qquad (1)$$

where k is the perturbation wavenumber, g is the ablator acceleration, L_ρ is the density gradient scale length of the ablation front, $v_{\rm a}$ is the ablation velocity, and α_2 and β_2 are parameters of order unity whose exact values depend on a heat conduction scale-length parameter, v, and the Froude number, ${\rm Fr} = v_{\rm a}^2/(gL_\rho)$. The key stabilizing effects of the high-foot drive enter through the higher ablation velocity, which scales as $T_{\rm rad}^{9/10}$, increasing the $\beta_2 kv_{\rm a}$ ablative stabilization term of equation (1), and through an increase in L_ρ, which reduces the instability term proportional to $\sqrt{kg}$. The increase in L_ρ is primarily due to a stronger first shock, which increases the adiabat of the implosion and prevents the ablator from becoming so highly compressed (risking break-up) during the implosion. The enhanced stability can be further understood by comparing the respective in-flight aspect ratios ($R_{\rm in}/\Delta R$, where $R_{\rm in}$ is the ablator inner radius and ΔR is the ablator thickness) of the high- and low-foot implosions: for the high-foot implosion, the in-flight aspect ratio is roughly half that of the low-foot implosion—the amplitude of instability growth is directly related to the exponential of $\sqrt{R_{\rm in}/\Delta R}/2$ (ref. 11). The trade-off made to obtain the improved stability of the high-foot implosion is that the D–T fuel adiabat, $a = P/P_{\rm F}$ (usually denoted α), where P is pressure and $P_{\rm F}$ is the Fermi pressure, is higher, making the fuel less compressible for a given amount of absorbed energy. (An alternate definition of the adiabat using $P_{\rm cold}$, the minimum D–T pressure at $1000\,{\rm g\,cm^{-3}}$, is sometimes used[12].) Details of the stability benefits, other theoretical motivations and trade-offs involved in the high-foot implosion, and the initial results from the first set of five D–T implosion experiments are described elsewhere[2,3].

Deuterium–tritium implosions N130927 and N131119 (NIF shot number in year–month–day format YYMMDD) build on the previous high-foot shot, N130812[3], by modestly increasing the NIF laser power and energy (Table 1) and by redistributing energy between different laser beams, through laser wavelength changes that affect the cross-beam transfer (the transfer of power from one beam to another via induced Brillouin scattering), to optimize the illumination pattern in

[1]Lawrence Livermore National Laboratory, PO Box 808, Livermore, California 94551, USA. [2]Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA.

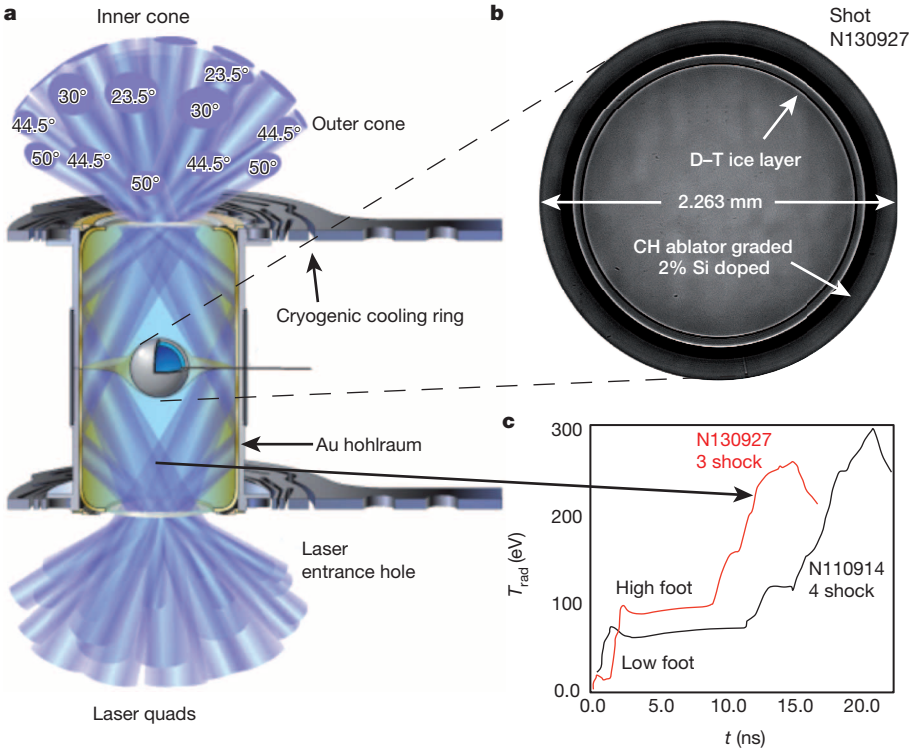


Figure 1 | Indirectly driven, inertially confined fusion target for NIF. **a**, Schematic NIF ignition target showing a cut-away of the gold hohlraum and plastic capsule with representative laser bundles incident on the inside surface of the hohlraum. **b**, X-ray image of the actual capsule for N130927 with D–T fuel layer and surrounding CH (carbon–hydrogen) plastic ablator. **c**, X-ray radiation drive temperature versus time for the NIC low-foot implosion and the post-NIC high-foot implosion.

the hohlraum[13–16]. Although the hotspot shape changes that result from these wavelength changes can be predicted to some extent[17], in practice the precise wavelengths needed to achieve the desired (that is, round) shape are found empirically. For N130927, the choice of $\lambda_{23.5} - \lambda_{30} = 0.7$ Å between the 23.5° and 30.0° inner-cone beams was chosen for azimuthal symmetry control, with $\Delta\lambda_{23.5-outer} = 9.2$ Å and $\Delta\lambda_{30-outer} = 8.5$ Å (the respective laser wavelength differences between the 23.5° and 30° inner-cone beams and the outer-cone beams) used for equatorial symmetry control (see Fig. 1 for beam angles). For N131119, $\Delta\lambda_{23.5-outer} = 9.5$ Å and $\Delta\lambda_{30-outer} = 8.8$ Å. These wavelength choices were critical for keeping the hotspot shape under control as the implosion was pushed to higher velocities, because previous experiments had already shown the tendency for the hotspot to deform into an oblate toroidal shape when laser power was increased[3]. There are limits to the amount of control that can be exerted over the hotspot shape just through wavelength changes alone, and physical changes to the hohlraum may also be required in future experiments to maintain hotspot (and fuel) shapes that will achieve the desired results.

We used a gold hohlraum of 5.75-mm diameter and 9.425-mm length, which are typical values in most high-foot cryogenic D–T implosion experiments (Fig. 1). The same hohlraum geometry was used during the NIC for most of the low-foot shots. As is typical for the high-foot series, the hohlraum was filled with helium gas of 1.6 mg cm^{-3} density (as compared with 0.96 mg cm^{-3} for the NIC), the purpose of which is to restrict and delay ingress of gold plasma from the inside wall of the hohlraum, which can impede laser beam propagation. The plastic capsule at the centre of the hohlraum for N130927 and N131119 respectively had outer-shell radii of 1.1315 and 1.1241 mm and inner-shell radii of 0.9365 and 0.9303 mm (Fig. 1). Layered on the inner surface of the capsule shell for N130927 and N131119 were 71.4

Table 1 | Measured and derived implosion performance metrics

Quantity	N131119$^{425\ TW}_{1.9\ MJ}$	N130927$^{390\ TW}_{1.8\ MJ}$	N130927[25]	N130927[26]	N130927 (sim.)
Y_{13-15} (neutron)	$(5.2 \pm 0.097) \times 10^{15}$	$(4.4 \pm 0.11) \times 10^{15}$	—	—	7.6×10^{15}
T_{ion} (keV) D–T	5.0 ± 0.2	4.63 ± 0.31	—	—	4.2
T_{ion} (keV) D–D	4.3 ± 0.2	3.77 ± 0.2	—	—	3.9
DSR (%)	4.0 ± 0.4	3.85 ± 0.41	—	—	4.1
τ_x (ps)	152.0 ± 33.0	161.0 ± 33.0	—	—	137
$P0_x$, $P0_n$ (μm)	35.8 ± 1.0, 34 ± 4	35.3 ± 1.1, 32 ± 4	—	—	32
$P2/P0_x$	−0.34 ± 0.039	−0.143 ± 0.044	—	—	—
$P3/P0_x$	0.015 ± 0.027	−0.004 ± 0.023	—	—	—
$P4/P0_x$	−0.009 ± 0.039	−0.05 ± 0.023	—	—	—
Y_{total} (neutron)	6.1×10^{15}	5.1×10^{15}	—	—	8.9×10^{15}
E_{fusion} (kJ)	17.3	14.4	—	—	25.1
r_{hs} (μm)	36.6	35.5	34.4–42.3	35.7–36.0	32.2
$(\rho r)_{hs}$ (g cm^{-2})	0.12–0.15	0.12–0.18	0.13–0.19	0.1–0.14	0.15
E_{hs} (kJ)	3.9–4.4	3.5–4.2	3.7–5.5	3.71–4.56	4.1
E_α (kJ)	2.2–2.6	2.0–2.4	2.0–2.4	2.0–2.5	2.8
$E_{DT,total}$ (kJ)	8.5–9.4	10.2–12.0	10.0–13.9	10.92–11.19	13.4
G_{fuel}	1.8–2.0	1.2–1.4	1.04–1.44	1.28–1.31	1.9

Lines 1–9 for columns 2 and 3 are directly measured quantities; others are derived from the data. Columns 4–6 show results from two data-driven models and simulation, respectively.

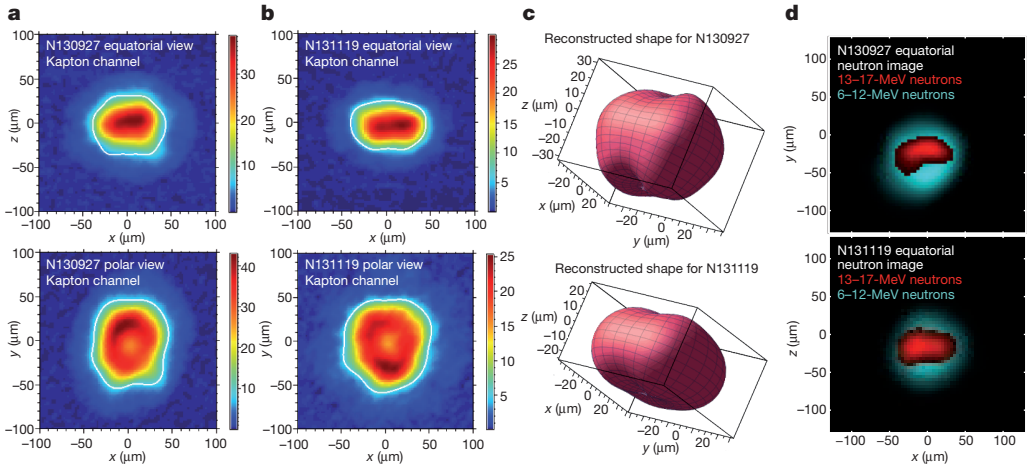


Figure 2 | X-ray and neutron images of the hotspot at bang-time. **a**, Equatorial (side-on) and polar (top-down) views of the hotspot shape for N130927. Kapton is the filter material in the imaging system that allows transmission of X-rays with energies of more than 6 keV. **b**, As in **a**, but for N131119. In these X-ray images, the contour shown in white is taken at the 17%-peak-brightness level (the colour scales show the brightness in arbitrary units) and is used to obtain a description of the shape in Legendre modes (equatorial view) and Fourier modes (polar view). **c**, Three-dimensional reconstructions of the hotspots. **d**, Superposition of direct (13–17 MeV) and down-scattered (6–12 MeV) neutron images from N130927 and N131119. (X-ray image analysis courtesy of N. Izumi, S. Khan, T. Ma and A. Pak of the NIF Shape Working Group; neutron image analysis courtesy of D. Fittinghoff, G. Grim, N. Guler and F. Merrill of the NIF Neutron Imaging System Working Group.)

and, respectively, 69.3 μm of cryogenic D–T ice that was held at 0.8 K below its triple point for a shot temperature of 18.6 K, like all high-foot D–T shots. This 'mini-quench condition' generally produces an ice layer with fewer ice cracks than that produced by a full quench[18]. Characterization of the capsule surface showed a roughness typical of implosion capsules for NIF, and characterization of the D–T ice showed a roughness well within requirements. The very high quality of the D–T ice layer on N130927 was probably not a significant factor in its performance because the third-best shot on the NIF (N130812) had an ice layer that was somewhat worse (with greater surface roughness) than average. The quality of the D–T layer for N131119 was between the qualities of the layers for N130812 and N130927.

Table 1 shows the key measurements and performance metrics for NIF shots N130927 and N131119. Key measured quantities are the neutron yield, Y_{13-15}, in the 13–15-MeV energy band around the characteristic 14.1-MeV D–T fusion neutron energy; the burn-averaged ion temperature, T_{ion}; the neutron and X-ray burn widths, respectively τ_n and τ_x; the down-scatter ratio (DSR); and the time of peak neutron brightness, or 'bang-time', t_b. On the NIF, Y_{13-15} is an average of many diagnostics, including four neutron time-of-flight (NToF) detectors[19], numerous radiochemical activation measurements[20] and a magnetic recoil spectrometer[21]. The temperature T_{ion} is directly related to the temporal spread obtained from the full-width at half-maximum of the NToF detectors. A temporal γ-ray history gives τ_n (for the high-foot experiments, τ_x and τ_n are consistent to within their errors). The DSR comes from measuring, via NToF and the magnetic recoil spectrometer, the number of neutrons scattered into the energy range 10–12 MeV, and is directly related to the areal density of the cold D–T fuel, $(\rho r)_{fuel} \approx 20.3f \times$ DSR (where f depends upon the amount of ablator mass remaining but is typically 0.95 ± 0.05 (ref. 21 and B. K. Spears, personal communication)). Other diagnostics such as X-ray imaging and neutron imaging (Fig. 2) give information on the shape of the implosion.

In what follows, we will use the aforementioned observables, which are measured over the duration of the fusion burn, to infer the amount of energy that was deposited into the D–T (both fuel and hotspot), to make a comparison with the amount of energy generated from fusion. The details of the analysis will focus on N130927; the results for N131119, which exceeded the performance of N130927, are quoted in Table 1. The analysis outlined in this letter uses an essentially one-dimensional 'onion-skin' picture with a hotspot of uniform density and temperature surrounded by the fuel (with Gaussian or uniform radial density profile), although the observed three-dimensional hotspot shape information is used to obtain the hotspot volume. Also, an assumption of approximately equal ion and electron temperatures, $T_{ion} \approx T_e$, is made and can be justified *post hoc* using an expression for the electron–ion collision time after the hotspot density is obtained. Analytical and simulation results based on less simplified assumptions are also quoted in Table 1 for comparison with what is detailed below.

By analysing the observed hotspot shape (Fig. 2) in terms of Legendre modes (equatorial view, lines 6–9 of Table 1) and Fourier modes (polar view), where the hotspot perimeter, as defined by the 17%-of-peak-brightness contour, is given by

$$R_{hs}(\theta) = P0\left[1 + \sum_{\ell=2}^{\infty}\left(\frac{P\ell}{P0}\right)P_\ell(\cos\theta)\right] \quad (2)$$

where $P_\ell(\cos\theta)$ is the Legendre function, we obtain the hotspot volume, V_{hs} (Methods), and the effective spherical radius, $r_{hs} = (3V_{hs}/4\pi)^{1/3}$. (We note that there is no absolute reference for the X-ray or neutron images, and so mode $\ell = 1$ is not included in the shape description. However, $\ell = 1$ and $m = 1$ motions can be obtained from the NToF detectors.) The total neutron yield, Y_{total}, can be calculated from $Y_{total} = Y_{13-15}\exp(4DSR)$, which accounts for the neutrons produced but then scattered out of the measured 13–15-MeV energy band by the cold and dense D–T fuel. Because for D–T fusion reactions the energy per fusion is known (14.1 MeV per neutron and 3.5 MeV per α-particle), E_{fusion}, the total fusion energy produced, can be calculated from Y_{total}.

From the measured T_{ion}, the D–T reaction rate per unit volume, $\langle\sigma v\rangle$, can be calculated using standard formulae[22] (Methods). For N130927, $\langle\sigma v\rangle = 4.75 \times 10^{-18}$–$1.03 \times 10^{-17}$ cm^3 s^{-1} The range of values is driven by the measurement uncertainty in T_{ion}. The reported T_{ion} values are actually averages over several detectors. The observed spread in the individual detector T_{ion} interpretations indicates some motional broadening contribution, which suggests that the lower temperature is more representative of the thermal temperature. Throughout this Letter, the uncertainty ranges given for values for all quoted quantities are driven by the uncertainty in T_{ion}.

For a 50:50 D–T mix the fusion power density is $\epsilon_{DT} = 7.04 \times 10^{-13} n^2 \langle\sigma v\rangle$ in joules per cubic centimetre per second, where n is the yet-unknown number density of the fusing region. From E_{fusion}, V_{hs} and

τ_x the hotspot number density can be calculated:

$$n = \sqrt{\frac{E_{\rm fusion}}{7.04\times10^{-13}\langle\sigma v\rangle V_{\rm hs}\tau_x}}$$

For N130927, $n = 8.1\times10^{24}$–$1.2\times10^{25}\,{\rm cm}^{-3}$, a value that also provides the hotspot mass density (assuming a pure D–T hotspot, with average atomic mass number $\bar{A}=2.5$ for D–T), $\rho_{\rm hs} = 34$–$50\,{\rm g\,cm}^{-3}$; the hotspot mass, $m_{\rm hs} = \rho_{\rm hs}V_{\rm hs} = 6.4$–$9.4\,\mu{\rm g}$; and the areal density, $(\rho r)_{\rm hs}$ (Table 1).

A number of quantities describing the implosion energetics now straightforwardly follow. The hotspot pressure can be obtained from $P_{\rm hs} = (\bar{Z}+1)\rho_{\rm hs}T_{\rm ion}/\bar{A}$ ($\bar{Z}=1$ for D–T), yielding $P_{\rm hs} = 126$–152 Gbar. The hotspot energy is then $E_{\rm hs} = (3/2)P_{\rm hs}V_{\rm hs}$ (Table 1). The fraction, f_α, of α-particle energy deposited into the hotspot can be calculated from a classic formula[23]

$$f_\alpha = 1 - \frac{1}{4\left[(\rho r)_{\rm hs}/\rho\lambda_\alpha\right]} + \frac{1}{160\left[(\rho r)_{\rm hs}/\rho\lambda_\alpha\right]^3}$$

where the α-particle stopping range can be found from[24]

$$\rho\lambda_\alpha = \frac{0.025T_{\rm e}^{5/4}}{1+0.0082T_{\rm e}^{5/4}} \qquad (3)$$

in base units of centimetres, grams, and kiloelectronvolts. For N130927, $f_\alpha = 0.68$–0.82. The energy deposited in the hotspot by α-particles is $E_\alpha = f_\alpha E_{\rm fusion}/5$, recalling that one-fifth of the D–T fusion energy is emitted in the form of α-particles (the remaining α-particle energy is deposited into the cold fuel). We note that, using the values found in Table 1, $E_\alpha/E_{\rm hs} \approx 0.56$. These energies fully describe the hotspot, but part of the implosion energy was used to compress the remaining cold D–T fuel and so we must examine the fuel to get a full picture of the implosion energy balance.

Because the D–T hotspot is formed by ablating the inner surface of the cold D–T fuel as electron conduction transports heat from the forming hotspot into the fuel, we can calculate the amount of D–T fuel remaining after the hotspot has formed because we know the initial amount of D–T ice layered onto the inside of the capsule, $m_0 = 186\,\mu{\rm g}$ (for N130927): $m_{\rm fuel} = m_0 - m_{\rm hs} = 176$–$179\,\mu{\rm g}$. The cold D–T fuel mass forms a shell surrounding the hotspot with volume $V_{\rm fuel} = (4\pi/3)\left(r_{\rm out}^3 - r_{\rm hs}^3\right)$, where $r_{\rm out}$ is the unknown outer fuel radius. Because $m_{\rm fuel} = 4\pi\int\rho_{\rm fuel}r^2\,{\rm d}r$ and the measured DSR provides a way to obtain the fuel density, $\rho_{\rm fuel}$, from $(\rho r)_{\rm fuel} = \int\rho_{\rm fuel}\,{\rm d}r$, by assuming a fuel profile we can solve for both the fuel layer thickness, $r_{\rm out} - r_{\rm hs}$, and density $\rho_{\rm fuel}$ (Methods). We find that

$$r_{\rm out} - r_{\rm hs} = 2\sigma = \sqrt{\frac{m_{\rm fuel}}{2\pi(\rho r)_{\rm fuel}} - r_{\rm hs}^2} - r_{\rm hs} \qquad (4)$$

with a Gaussian density profile

$$\rho(r) = (\rho r)_{\rm fuel}\frac{\exp\left[-(r-r_0)^2/2\sigma^2\right]}{\sqrt{2\pi}\sigma}$$

where r_0 is the radius of peak fuel density. For N130927, $r_{\rm out} - r_{\rm hs} = 14.7$–$15.3\,\mu{\rm m}$, $\rho_{\rm fuel} = 385$–$402\,{\rm g\,cm}^{-3}$ and $V_{\rm fuel} = (3.0$–$3.2)\times10^5\,\mu{\rm m}^3$. The fuel outer radius from these arguments, $r_{\rm out} = 50.8\,\mu{\rm m}$ (at 50% $\rho_{\rm fuel}$), is close to that obtained directly from the down-scattered neutron image (Fig. 2), where $P0 = 55.4\,\mu{\rm m}$ (at 17% of maximum intensity). By this time of peak compression, the D–T fuel density has increased by a factor of more than 1,500. The fuel density is not required for calculating the fuel energy, but it can be used to estimate the adiabat of the fuel (at bang-time) assuming that the cold fuel and the hotspot are isobaric ($P_{\rm fuel} \approx P_{\rm hs}$), in which case we find that $a = P_{\rm fuel}/P_{\rm F} \approx P_{\rm hs}/0.0021\rho_{\rm fuel}^{5/3} = 2.9$–$3.3$ for N130927—the fuel adiabat in flight is lower than this range of values. The fuel density is also needed to calculate the X-ray losses through the fuel.

As the hotspot is compressed to high temperatures, the primary energy loss mechanism is bremsstrahlung X-ray emission because the D–T hotspot is optically thin to these X-rays. The bremsstrahlung energy loss is calculated to be[24]

$$E_{\rm brems}\ ({\rm kJ}) = 5.34\times10^{-34}n_{\rm hs}^2\sqrt{T_{\rm e}}V_{\rm hs}\tau_x$$

in base units of centimetres, kiloelectronvolts and seconds. For N130927, $E_{\rm brems} = 2.3$–$4.5\,{\rm kJ}$, the low end of which is nearly equivalent to the

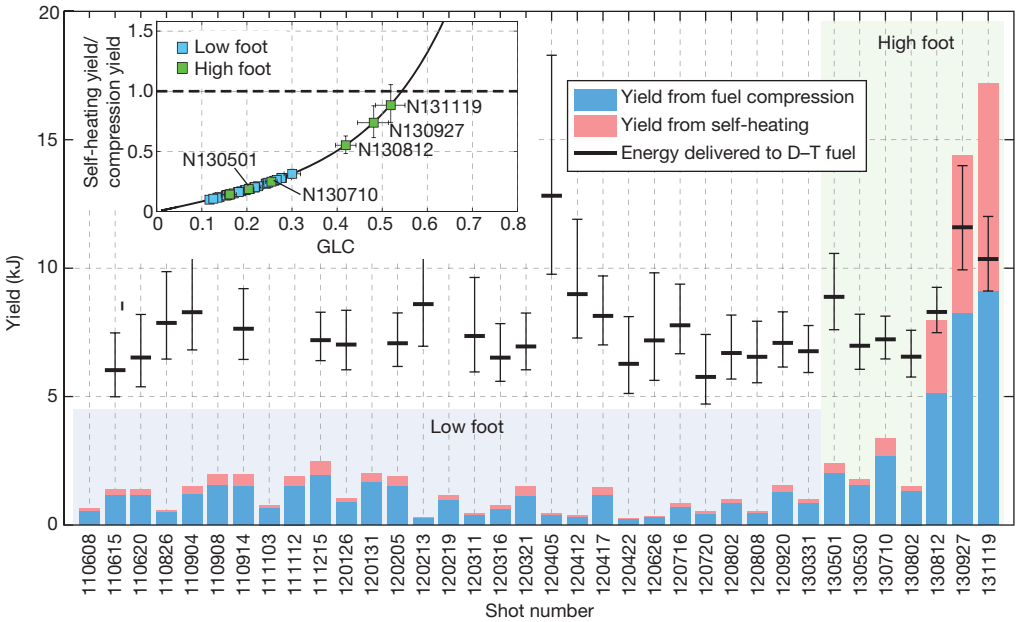


Figure 3 | Yield and energetics metrics for shots on the NIF. Total fusion yield is plotted versus shot number (that is, time). Shots 110608–130331 are low-foot shots. Shots 130501–131119 are high-foot shots. The bars showing total yield are broken into components of yield coming from α-particle self-heating and yield coming from compression. The black dashes denote the energy delivered to the D–T (fuel plus hotspot) with error bars (black vertical lines, 1σ) as calculated from the model of ref. 25. The plot shows that, even with the uncertainty in our results, shots 130927 and 131119 both yielded more fusion energy than was delivered to the D–T. Inset, ratio of self-heating yield to compression yield versus generalized Lawson criterion (GLC). All error bars, 1σ.

α-particle energy deposited. To examine whether or not these X-rays can escape the dense cold fuel, we can calculate the optical depth of the cold D–T fuel from $\tau_{\rm fuel} = \rho_{\rm fuel}\kappa_{\rm DT}(r_{\rm out} - r_{\rm hs})$ using a simple D–T opacity model, $\kappa_{\rm DT}\,({\rm g\,cm^{-2}}) = 0.352\rho_{\rm fuel}(h\nu)^{-3.3}[1 - \exp(-h\nu/T_{\rm e})]$. We find that for X-ray energies of $h\nu \approx T_{\rm ion} \approx T_{\rm e}$, the D–T fuel layer is almost one optical depth, $\tau_{\rm fuel} = 0.32$–0.66, implying that some bremsstrahlung X-rays deposit energy into the cold fuel whereas some escape. Electron conduction does not have a significant role in the total D–T energy loss from the cold fuel at stagnation, but is important for the hotspot energy loss.

The cold fuel energy at stagnation now follows from the isobaric assumption, $E_{\rm fuel} = (3/2)P_{\rm fuel}V_{\rm fuel} = 6.9$–$7.8\,{\rm kJ}$ (where we have overestimated the fuel internal energy because typically the outer edge of the fuel has not fully stagnated even at bang-time). The total energy delivered to the D–T by the implosion is then (Table 1)

$$E_{\rm DT,total} = E_{\rm hs} + E_{\rm fuel} + \frac{1}{2}{\rm e}^{-\tau_{\rm fuel}}E_{\rm brems} - \frac{1}{2}E_{\alpha} \qquad (5)$$

The factors of one-half in the radiation term and the α-particle energy deposition account for having only half the energy emitted or deposited at peak burn. This total D–T energy was calculated with quantities measured around bang-time, but it represents the kinetic plus internal energy in the fuel at peak velocity in the implosion. A crosscheck of $E_{\rm DT,total}$ is provided from calculating the fuel kinetic energy, K, using a direct measurement of implosion velocity from an earlier high-foot '1DConA' shot, N130409 (at 350 TW and 1.3 MJ of laser power and energy), where the peak ablator centre-of-mass velocity was measured to be $267 \pm 15\,{\rm km\,s^{-1}}$, which is equivalent to a fuel velocity of $297 \pm 15\,{\rm km\,s^{-1}}$ (the fuel being at smaller radius and convergence makes the velocity larger). Scaling the N130409-derived velocity to the laser power of N130927 (implosion velocity, $v_{\rm imp} \propto P_{\rm laser}^{0.41}$) gives a fuel velocity of $v_{\rm fuel} = 311 \pm 15\,{\rm km\,s^{-1}}$, and so $K = (1/2)m_0 v_{\rm fuel}^2 = 9.0 \pm 0.9$ kJ. The difference between K and $E_{\rm DT,total}$ is the internal energy of the fuel at peak velocity plus the additional PdV work done by the ablator on the fuel during the deceleration.

The total fuel energy gain, $G_{\rm fuel} = E_{\rm fusion}/E_{\rm DT,total}$ is now known and is 1.2–1.4 for N130927. For comparison, in Table 1 we also show results from other data-derived models of implosion energetics[25,26] that are constructed in the spirit of the above analysis but which differ in some details. A conduction-limited temperature profile in the hotspot is added to the above development in one case[25] and the other 'detailed model' case includes a three-dimensional, self-consistent physics model matched to the data[26]. To complement these analytic data-driven models, in Table 1 we also show the results from a full, one-dimensional, radiation–hydrodynamics simulation[27] of N130927, with a multifrequency X-ray drive that is calibrated to shock-timing and implosion trajectory data, without any mix model applied. The inferences from data and the computer simulation all indicate that $G_{\rm fuel} > 1$. Moreover, we have demonstrated repeatability and improvement with the follow-on shot N131119. It should be understood, however, that $G_{\rm fuel} > 1$ indicates only that the output fusion energy exceeds the energy deposited into the fuel. This is not the same as exceeding either the energy absorbed by the capsule (defined as the ablator shell plus D–T fuel), which absorbed ~150 kJ for N130927 and N131119, or the energy delivered by the laser to the target (defined as the hohlraum plus capsule), which was 1.8 MJ for N130927 and 1.9 MJ for N131119.

Key yield and energy performance metrics are graphically illustrated in Fig. 3 for N131119, N130927 and all other D–T implosions carried out on the NIF since the summer of 2011. Using a key metric for ignition, the generalized Lawson criterion[28] $\chi = (P\tau)/(P\tau)_{\rm ign}$ (which is unity at ignition) we see (Fig. 3, inset) that for N131119 we are at the threshold of achieving yield doubling due to α-particle energy deposition.

Because most of the quantities associated with inertially confined fusion that we seek to improve to achieve ignition scale as some positive power of stagnation pressure, near-term efforts focus on increasing the implosion speed and controlling the hotspot shape with the present fuel adiabat. As the implosion speed is increased, we will necessarily risk giving back some of the gains the high-foot implosion has made in terms of instability control. New strategies for the hohlraum will also be explored because at present hohlraum physics is limiting our ability to use the full power capability on the NIF while maintaining an acceptable hotspot shape (higher laser powers are the most direct way to increase implosion speed). Future efforts may involve more elaborate schemes to maintain control over ablator instability while recovering a lower adiabat for the fuel (for example 'adiabat shaping'[29]) or also using an alternate ablator material.

METHODS SUMMARY

Formulae for the hotspot volume and the D–T reaction rate, and a discussion of fuel density profiles, are given in Methods. Neutron image shape coefficients are also given there.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 November 2013; accepted 7 January 2014.
Published online 12 February 2014.

1. Edwards, M. J. *et al.* Progress towards ignition on the National Ignition Facility. *Phys. Plasmas* **20,** 070501 (2013).
2. Dittrich, T. R. *et al.* Design of a high-foot/high-adiabat ICF capsule for the National Ignition Facility. *Phys. Rev. Lett.* **112,** LJ14108 (2014).
3. Park, H.-S. *et al.* High-adiabat, high-foot, inertial confinement fusion implosion experiments on the National Ignition Facility. *Phys. Rev. Lett.* **112,** LK13998 (2014).
4. Lindl, J. D. & Moses, E. I. Plans for the National Ignition Campaign (NIC) on the National Ignition Facility (NIF): on the threshold of initiating ignition experiments. *Phys. Plasmas* **18,** 050901 (2011).
5. Glenzer, S. H. *et al.* Cryogenic thermonuclear fuel implosions on the National Ignition Facility. *Phys. Plasmas* **19,** 056318 (2012).
6. Bodner, S. E. Rayleigh-Taylor instability and laser-pellet fusion. *Phys. Rev. Lett.* **33,** 761–764 (1974).
7. Gonchorov, V. N. & Hurricane, O. A. *Panel 3 Report: Implosion Hydrodynamics.* Report LLNL-TR-562104 (Lawrence Livermore National Laboratory, 2012).
8. Ma, T. *et al.* Onset of hydrodynamic mix in high-velocity, highly compressed inertial confinement fusion implosions. *Phys. Rev. Lett.* **111,** 085004 (2013).
9. Regan, S. P. *et al.* Hot-spot mix in ignition-scale inertial confinement fusion targets. *Phys. Rev. Lett.* **111,** 045001 (2013).
10. Betti, R., Goncharov, V. N., McCrory, R. L. & Verdon, C. P. Growth rates of the Rayleigh-Taylor instability in inertial confinement fusion. *Phys. Plasmas* **5,** 1446–1454 (1998).
11. Lindl, J. Development of the indirect-drive approach to inertial confinement fusion and the target physics basis for ignition and gain. *Phys. Plasmas* **2,** 3933–4024 (1995).
12. Haan, S. *et al.* Point design targets, specifications, and requirements for the 2010 ignition campaign on the National Ignition Facility. *Phys. Plasmas* **18,** 051001 (2011).
13. Michel, P. *et al.* Tuning the implosion symmetry of ICF targets via controlled crossed-beam energy transfer. *Phys. Rev. Lett.* **102,** 025004 (2009).
14. Michel, P. *et al.* Symmetry tuning via controlled crossed-beam energy transfer on the National Ignition Facility. *Phys. Plasmas* **17,** 056305 (2010).
15. Moody, J. D. *et al.* Multistep redirection by cross-beam power transfer of ultrahigh-power lasers in a plasma. *Nature Phys.* **8,** 344–349 (2012).
16. Callahan, D. A. *et al.* The velocity campaign for ignition on NIF. *Phys. Plasmas* **19,** 056305 (2012).
17. Marinak, M. M. *et al.* Three-dimensional HYDRA simulations of National Ignition Facility targets. *Phys. Plasmas* **8,** 2275 (2001).
18. Kozioziemski, B. J. *et al.* Deuterium-tritium layer formation for the National Ignition Facility. *Fusion Sci. Technol.* **59,** 14–25 (2011).
19. Glebov, V. Yu. *et al.* Development of nuclear diagnostics for the National Ignition Facility. *Rev. Sci. Instrum.* **77,** 10E715 (2006).
20. Bleuel, D. L. *et al.* Neutron activation diagnostics at the National Ignition Facility. *Rev. Sci. Instrum.* **83,** 10D313 (2012).
21. Gatu Johnson, M. *et al.* Neutron spectrometry – an essential tool for diagnosing implosions at the National Ignition Facility. *Rev. Sci. Instrum.* **83,** 10D308 (2012).
22. Bosch, H.-S. & Hale, G. M. Improved formulas for fusion cross-section and thermal reactivities. *Nucl. Fusion* **32,** 611–631 (1992).
23. Krokhin, O. N. & Rozanov, V. B. Escape of α-particles from a laser-pulse-initiated thermonuclear reaction. *Sov. J. Quantum Electron.* **2,** 393–394 (1973).
24. Atzeni, S. & Meyer-ter-Vehn, J. *The Physics of Inertial Fusion* 32, 398 (Oxford Univ. Press, 2004).
25. Patel, P. *et al.* Performance of DT layered implosions on the NIF. *Bull. Am. Phys. Soc.* **58,** abstr. NO4.00001 (2013).

26. Cerjan, C., Springer, P. T. & Sepke, S. M. Integrated diagnostic analysis of inertial confinement fusion capsule performance. *Phys. Plasmas* **20,** 056319 (2013).
27. Zimmerman, G. B. & Kruer, W. L. Numerical Simulation of laser initiated fusion. *Comments Plasma Phys. Control. Fusion* **2,** 85–89 (1975).
28. Betti, R. *et al.* Thermonuclear ignition in inertial confinement fusion and comparison with magnetic confinement. *Phys. Plasmas* **17,** 058102 (2010).
29. Goncharov, V. N. *et al.* Improved performance of direct-drive ICF target designs with adiabat shaping using an intense picket. *Phys. Plasmas* **10,** 1906–1918 (2003).

Acknowledgements We thank P. Albright, J. Atherton, L. R. Benedetti, D. Bradley, J. A. Caggiano, R. Dylla-Spears, M. J. Edwards, W. H. Goldstein, B. Goodwin, S. Haan, A. Hamza, W. Hsing, P. Kervin, J. Kilkenny, B. Kozioziemski, O. Landen, J. Lindl, B. MacGowan, A. Mackinnon, N. Meezan, J. F. Meeker, J. Moody, E. Moses, D. Pilkington, T. Parham, J. Ralph, S. Ross, H. Robey, R. Rygg, B. Spears, R. Town, C. Verdon, A. Wan and B. Van Wonterghem, and the NIF operations, cryogenics and targets teams. We also thank V. Goncharov and J. Knauer for their advice, and R. Betti for bringing our attention to equation (3). Thanks also go to NIF's external collaborators at GA (targets), LLE (diagnostics), the MIT Plasma Science and Fusion Center (magnetic recoil spectrometer diagnostic), CEA and AWE. This work was performed under the auspices of the US Department of Energy by Lawrence Livermore National Laboratory under contract no. DE-AC52-07NA27344.

Author Contributions O.A.H. was lead scientist for the high-foot campaign, and performed two-dimensional stability modelling, and one-dimensional pre- and post-shot analysis. D.A.C. was lead scientist on hotspot shape and hohlraum strategies. D.T.C. was part of the D–T shot experiment team. P.M.C. performed VISAR data unfolds. C.C. performed three-dimensional 'detailed model' calculations. E.L.D. was lead experimenter for 1DConA ($R(t)$ trajectory) tuning experiments and capsule re-emission (early-time symmetry) tuning experiments. T.R.D. performed initial one-dimensional capsule design, scoping, and one-dimensional pre- and post-shot simulations. T.D. was lead experimentalist on a 2DConA ablator shape experiment and was part of the D–T shot team. D.E.H. was the pulse shape design physicist and performed all two-dimensional integrated hohlraum-capsule simulations. L.F.B.H. was design physicist for keyhole (shock-timing) tuning experiments. J.L.K. was lead experimentalist for symcap (hotspot shape) tuning experiments. S.L. was lead experimentalist for the keyhole (shock-timing) tuning experiments. T.M. was lead experimentalist for several 2DConA ablator shape experiments, was part of the D–T shot team, and was lead experimentalist on shot N131119. A.G.M. was part of the 1DConA and D–T experimental teams. J.L.M. was design physicist for the re-emission experiment. A.P. was part of the D–T shot team. P.K.P. provided a hotspot model analysis and metrics plots. H.-S.P. was lead experimentalist on D–T implosion shots up to and including N130927. B.A.R. was overall lead on experiments. J.D.S. constructed model multifrequency sources normalized to tuning experiments, and performed one- and two-dimensional model scoping. P.T.S. provided a hotspot model analysis. R.T. provided 1DConA analysis and was shot experimentalist.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to O.A.H. (hurricane1@llnl.gov).

METHODS

Hotspot volume formulae. Using equation (2) the volume is

$$V_{\rm hs} = 2\pi \int_0^{\pi} \int_0^{R_{\rm hs}(\theta)} R^2 \, {\rm d}R \sin\theta \, {\rm d}\theta$$
$$= \frac{4}{3}\pi P0^3 + \frac{4}{5}\pi P0P2^2 + \frac{8}{105}\pi P2^3 + \frac{4}{7}\pi P0P3^2$$
$$+ \frac{16}{105}\pi P2P3^2 + \frac{4}{9}\pi P0P4^2 + \frac{8}{35}\pi P2^2P4$$
$$+ \frac{80}{693}\pi P2P4^2 + \frac{8}{77}\pi P3^2P4 + \frac{24}{1{,}001}\pi P4^3 + \cdots$$

A simple correction to this volume can be applied (by multiplying the above expression by the expression below) to include m-modes (azimuthal modes):

$$1 + \frac{3}{2}\left(\frac{M2}{M0}\right)^2 + \frac{3}{2}\left(\frac{M3}{M0}\right)^2 + \frac{3}{2}\left(\frac{M4}{M0}\right)^2$$
$$+ \frac{3}{4}\frac{M2^2M4}{M0^3}\cos[4(\phi_2 - \phi_4)] \cdots$$

where ϕ_2 and ϕ_4 are phase offsets of m-modes 2 and 4, respectively.

Deuterium–tritium reaction rate formulae. From ref. 22

$$\langle \sigma v \rangle = C_1 \zeta^{-5/6} \xi^2 \exp\left(-3\zeta^{1/3}\xi\right)$$

in cubic centimetres per second, where

$$\xi = \frac{C_0}{T_{\rm ion}^{1/3}}$$

$$\zeta = 1 - \frac{C_2 T_{\rm ion} + C_4 T_{\rm ion}^2 + C_6 T_{\rm ion}^3}{1 + C_3 T_{\rm ion} + C_5 T_{\rm ion}^2 + C_7 T_{\rm ion}^3}$$

and $C_0 = 6.6610$, $C_1 = 6.4341 \times 10^{-14}$, $C_2 = 1.5136 \times 10^{-2}$, $C_3 = 7.5189 \times 10^{-2}$, $C_4 = 4.6064 \times 10^{-3}$, $C_5 = 1.35 \times 10^{-2}$, $C_6 = -1.0675 \times 10^{-4}$ and $C_7 = 1.366 \times 10^{-5}$ when $T_{\rm ion}$ is expressed in kiloelectronvolts.

Deuterium–tritium fuel density profile. Assuming a different density profile changes the form of equation (4), but changes the numerical value for the fuel thickness little. For example, assuming a top-hat distribution for the fuel yields

$$r_{\rm out} - r_{\rm hs} = \frac{1}{2}\sqrt{3m_{\rm fuel}/\pi(\rho r)_{\rm fuel} - 3r_{\rm hs}^2} - 3r_{\rm hs}/2$$

from which we obtain $r_{\rm out} - r_{\rm in} = 15.5$–$16.2\,\mu$m for N130927. The fuel density does show more sensitivity, being $\rho_{\rm fuel} = (\rho r)_{\rm fuel}/(r_{\rm out} - r_{\rm in}) = 457$–$478\,{\rm g\,cm^{-3}}$ for a top-hat distribution and $\rho_{\rm fuel} = (\rho r)_{\rm fuel}/\sqrt{2\pi}\sigma = 385$–$402$ for a Gaussian. The Gaussian profile assumption is more consistent with simulated fuel density profiles. The lower fuel density associated with the Gaussian profile increases the inferred fuel adiabat and decreases the fuel optical depth as compared with the uniform profile. The lower fuel optical depth makes the X-ray energy contribution to equation (5) larger; that is, it gives us a more conservative contribution to the total D–T energy.

Neutron image shape analysis. For N130927, the Legendre mode shape coefficients for the down-scattered neutron image are $P0 = 55 \pm 4\,\mu$m, $P2/P0 = 1\% \pm 5\%$ and $P4/P0 = -2\%$, and for the direct image $P0 = 32 \pm 4\,\mu$m, $P2/P0 = -35\% \pm 5\%$ and $P4/P0 = 2\%$. For the N131119 down-scattered neutron image, $P0 = 50 \pm 4\,\mu$m, $P2/P0 = 0\% \pm 5\%$ and $P4/P0 = 2\%$, and for the direct image $P0 = 34 \pm 4\,\mu$m, $P2/P0 = -34\% \pm 5\%$ and $P4/P0 = 1\%$.

LETTER

doi:10.1038/nature12952

Exceptional ballistic transport in epitaxial graphene nanoribbons

Jens Baringhaus[1]*, Ming Ruan[2]*, Frederik Edler[1], Antonio Tejeda[3,4], Muriel Sicot[3], AminaTaleb-Ibrahimi[4], An-Ping Li[5], Zhigang Jiang[2], Edward H. Conrad[2], Claire Berger[2,6], Christoph Tegenkamp[1] & Walt A. de Heer[2]

Graphene nanoribbons will be essential components in future graphene nanoelectronics[1]. However, in typical nanoribbons produced from lithographically patterned exfoliated graphene, the charge carriers travel only about ten nanometres between scattering events, resulting in minimum sheet resistances of about one kilohm per square[2–5]. Here we show that 40-nanometre-wide graphene nanoribbons epitaxially grown on silicon carbide[6,7] are single-channel room-temperature ballistic conductors on a length scale greater than ten micrometres, which is similar to the performance of metallic carbon nanotubes. This is equivalent to sheet resistances below 1 ohm per square, surpassing theoretical predictions for perfect graphene[8] by at least an order of magnitude. In neutral graphene ribbons, we show that transport is dominated by two modes. One is ballistic and temperature independent; the other is thermally activated. Transport is protected from back-scattering, possibly reflecting ground-state properties of neutral graphene. At room temperature, the resistance of both modes is found to increase abruptly at a particular length—the ballistic mode at 16 micrometres and the other at 160 nanometres. Our epitaxial graphene nanoribbons will be important not only in fundamental science, but also—because they can be readily produced in thousands—in advanced nanoelectronics, which can make use of their room-temperature ballistic transport properties.

The energy spectrum of a graphene ribbon with length L and width W is approximately given by

$$E_{n,m} = \pm \hbar c^* \sqrt{\left(\frac{n\pi}{W}\right)^2 + \left(\frac{m\pi}{L}\right)^2} \quad (1)$$

where $c^* \approx 10^6\ \mathrm{m\,s^{-1}}$ is the Fermi velocity and $\hbar$ is Planck's constant divided by 2π. For reference, if $W = 40$ nm and $L = 1$ μm, then $E_{1,0}/k_B = 600$ K and $E_{0,1}/k_B = 23$ K, where k_B is Boltzmann's constant[9]. Following the Landauer equation[10], the conductance G of a long graphene ribbon, measured in a two-probe measurement, is $G = 4G_0\sum \mathrm{Tr}_n$ where $G_0 = 1/R_0 = e^2/h$, and $0 \le \mathrm{Tr}_n \le 1$ is the corresponding transmission coefficient (e is the electronic charge). At low temperatures for $|E_{n,0}| \le |E_F|$, $\mathrm{Tr}_n \approx (1 + L/\lambda_n)^{-1}$, where λ_n is the mean free path and E_F is the Fermi energy[10]. For $|E_{n,0}| > |E_F|$, $\mathrm{Tr}_n = 0$. The $n = 0$ modes are special, and relate to edge states[11,12]. They dominate transport when $|E_F| < |E_{1,0}|$, that is, for charge-neutral ribbons with temperature $T <$ 600 K. By analogy with high-quality carbon nanotubes[13], charge-neutral low-defect graphene ribbons were expected to be micrometre-scale ballistic conductors[1]. However, in lithographically patterned exfoliated graphene ribbons, transport in the edge states is quenched due to disorder[2–5] (Supplementary Fig. 2).

In contrast, well-aligned, single-crystal monolayer graphene sheets form spontaneously on silicon carbide (SiC) surfaces heated above 1,000 °C. In the structured growth method[6], graphene ribbons self-assemble on the sidewalls of steps that are etched into the (0001) surface of electronics-grade SiC wafers[6,7,14–16] (see Figs 1a and 2), so that no graphene patterning is required to produce nanosized ribbons. More precisely, to prepare the ribbons shown below, 20-nm-deep trenches were etched along the SiC [1$\bar{1}$00] direction. The samples in Fig. 2 were annealed at 1,600 °C for 15 min (ref. 7). The samples in Fig. 3 were heated at 1,300 °C in Ar (4×10^{-5} mbar), and then in ultrahigh vacuum (UHV) for 15 min at 1,100 °C (green dots, Fig. 3a) or at 1,150 °C (all others—see Supplementary Information). The natural-step ribbon[7]

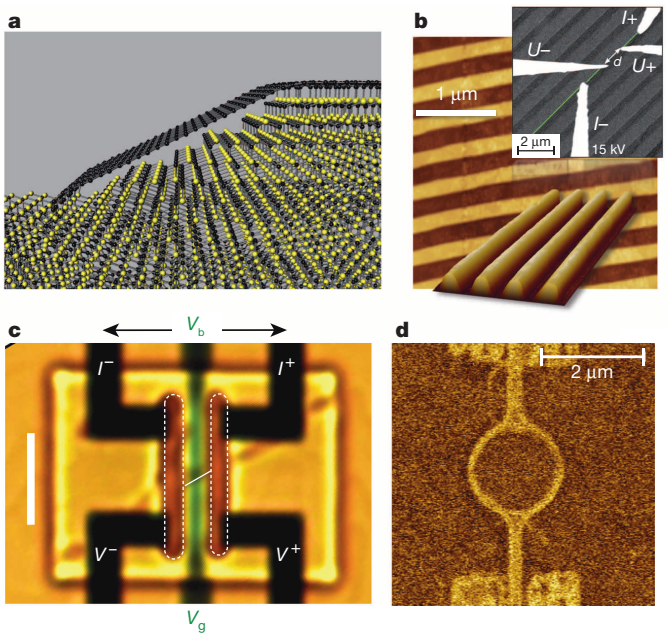


Figure 1 | Structure and characterization of nanoribbons and devices. **a**, Schematic diagram of a graphene ribbon (all black) on an annealed and faceted sidewall (yellow and black; see ref. 17). **b**, AFM image of an array of graphene ribbons on sidewalls of 20-nm-deep trenches. Bottom inset, three-dimensional view. Top inset, guided by an SEM, up to four individual probes are brought into contact with a selected graphene ribbon, and serve as current leads ($I+$, $I-$) and voltage probes ($V+$, $V-$). The sample can be transferred to and from an *in situ* heating stage for annealing up to 1,500 °C. **c**, Optical micrograph of sidewall ribbon (sample A) supplied with leads and gate consisting of wide graphene ribbons (1 μm apart) connected by a nominally 39-nm ribbon to form an H-shaped geometry, where the vertical, wide graphene ribbons serve as current leads (I^+, I^-) and voltage probes (V^+, V^-) for the 1.6-μm-long ribbon. White dashed lines indicate location of the graphene leads, white line indicates graphene ribbon. Green region locates the gate structure. Dark areas are the gold contacts. **d**, Electrostatic force image of a sidewall graphene nano-ring with 1.6-μm outer diameter attached to graphene leads. The ring is produced similarly to **c** and has graphene-covered sloping sidewalls.

[1]Institut für Festkörperphysik, Leibniz Universität, Hannover, Appelstrasse 2, 30167 Hannover, Germany. [2]School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332-0430, USA. [3]Université de Lorraine, UMR CNRS 7198, Institut Jean Lamour, BP 70239, 54506 Vandoeuvre-lès-Nancy, France. [4]UR1 CNRS/Synchrotron SOLEIL, Saint-Aubin, 91192 Gif sur Yvette, France. [5]Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Tennessee 37831, USA. [6]Institut Néel, CNRS UJF-INP, 38042 Cedex 6, Grenoble, France.
*These authors contributed equally to this work.

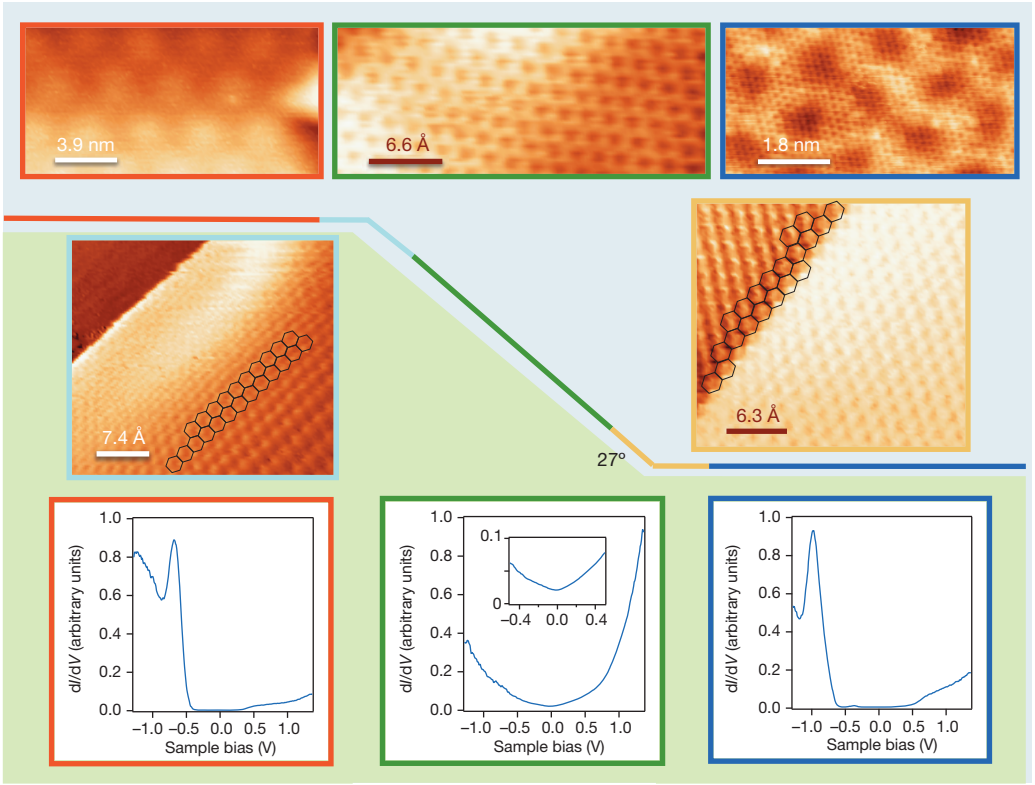


Figure 2 | Scanning tunnelling analysis of *ex situ* produced sidewall ribbons similar to those used in fixed geometry transport measurements. Colour-coded line over the 27° slope indicates areas of the surface investigated: colours correspond to the frame colours of the images. Top row: green frame shows atomic resolution STM of graphene structure on the sloped sidewall, and corresponds to the typical graphene STS traces shown in the bottom row (green frame); red frame shows STM of upper terrace, blue frame shows STM of lower terrace. Middle row: blue frame shows STM of upper edge, and yellow frame STM of lower edge; both frames show helical edge structures. Bottom row: red and blue frames show STS of upper and lower terraces respectively, and show semiconducting gap.

(Figs 1c and 4) was connected at each end to two 200-nm-deep graphitized trenches 1 μm apart. These sidewall ribbons have been extensively characterized[6,7,14–17] (Figs 1 and 2 and Supplementary Information) with scanning probe microscopies to determine ribbon widths and sidewall slopes. Angle-resolved photoemission spectroscopy (ARPES)[16] performed at the Cassiopée beam line at the Soleil synchrotron shows a Dirac cone (Supplementary Fig. 1), demonstrating that the ribbons are well-aligned monolayers, and that the sidewall slope is uniform (that is, about 28°, consistent with the $(2\bar{2}07)$ facet). Scanning tunnelling microscopy and scanning tunnelling spectroscopy (STM and STS) show that essentially charge-neutral graphene covers the sidewalls (Fig. 2). The top and bottom plateaus show the semiconducting properties of the 'buffer layer' (Fig. 2, red frames). Atomic-resolution images show zig-zag and chiral edges of the ribbons (Fig. 2).

Single-channel ballistic transport in epitaxial graphene ribbons was first reported[7,15] on 40-nm-wide natural-step ribbons that were seamlessly connected to wide graphene leads (see Fig. 1c and Supplementary Information). These ribbons, used for the measurement presented in Fig. 4, were supplied with a top gate (20-nm Al_2O_3 coated with aluminium) so that E_F could be adjusted. Four wires were bonded to the graphene leads, facilitating four-point transport measurements.

Subsequently, we performed *in situ* variable-geometry transport measurements (at temperatures T from 30 to 300 K) on ~40-nm-wide ribbons (Fig. 1b), confirming single-channel ballistic transport as discussed below. Four nanoscopically sharp tungsten probes were positioned using a built-in scanning electron microscope and brought into ohmic electrical contact with a selected ribbon (Fig. 1b) in UHV (Omicron Nanoprobe UHV system). For two-probe (2p) measurements, a current I_{12} was passed between two probes, in contact with the ribbon, and the voltage V_{12} was measured so that $R_{2p} = V_{12}/I_{12}$. For four-probe measurements, a current I_{34} was passed through outer probes and the potential difference V_{12} between the two inner probes (separated by a distance L) was measured: $R_{4p} = V_{12}/I_{34}$. The resistance per unit length R' is found from the slope of R_{4p} versus L for 1 μm ≤ L ≤ 16 μm (Fig. 3a). For these ribbons, R' ranges from 0 to 6 kΩ μm^{-1} (Fig. 1b). The R' values decrease after *in situ* heating (a process known to clean graphene), indicating that surface contamination increases scattering.

For these ribbons, both R_{4p} and R_{2p} extrapolate to $R \approx R_0 = h/e^2$ at $L = 0$ (Fig. 3a). Moreover, the resistivity ratio RR = $R_{4p}/R_{2p} = 0.95 \pm 0.02$. These results are characteristic of single-channel ballistic transport involving nearly perfectly invasive contacts[10,18,19]. Furthermore, because an invasive probe acts as a scattering centre, R_{2p} is strongly modified if a third passive probe (with a very large resistance to ground) is placed between the contact probes[10]. Left-moving charges that enter the passive probe from the left, say, will leave the probe going either left or right with equal probability. Hence the transmission probability is Tr = 0.5 so that the overall wire resistance R_{2p} doubles from R_0 to $2R_0$ (see Supplementary Information). When two passive probes are used, the overall resistance R_{2p} increases to $3R_0$. From these considerations it follows that RR = 1 for perfectly invasive probes[10,18,19]. This property of invasive probes is explicitly demonstrated in Fig. 3b. When one probe is placed on a ribbon, the end-to-end resistance R_{2p} increases from approximately R_0 to $2R_0$; when two probes are used, it increases to $3R_0$. Note that placing probes on a diffusive wire has no effect on the resistance of the wire. As shown in Fig. 3c, the resistance of these single-channel ballistic ribbons does not significantly depend on temperature or on the bias voltage.

For 0.1 μm ≤ L ≤ 1 μm, $R(L)$ is found to increase nonlinearly from $0.5R_0$ to $1R_0$, as shown for two ribbons (Fig. 3a, upper left inset). For $L < L^*_{0-} = 160$ nm, $R \approx R_0/2$, whereas for $L > L^*_{0-}$, R increases (see Fig. 3a upper left inset, and Supplementary Fig. 7). For the two ribbons,

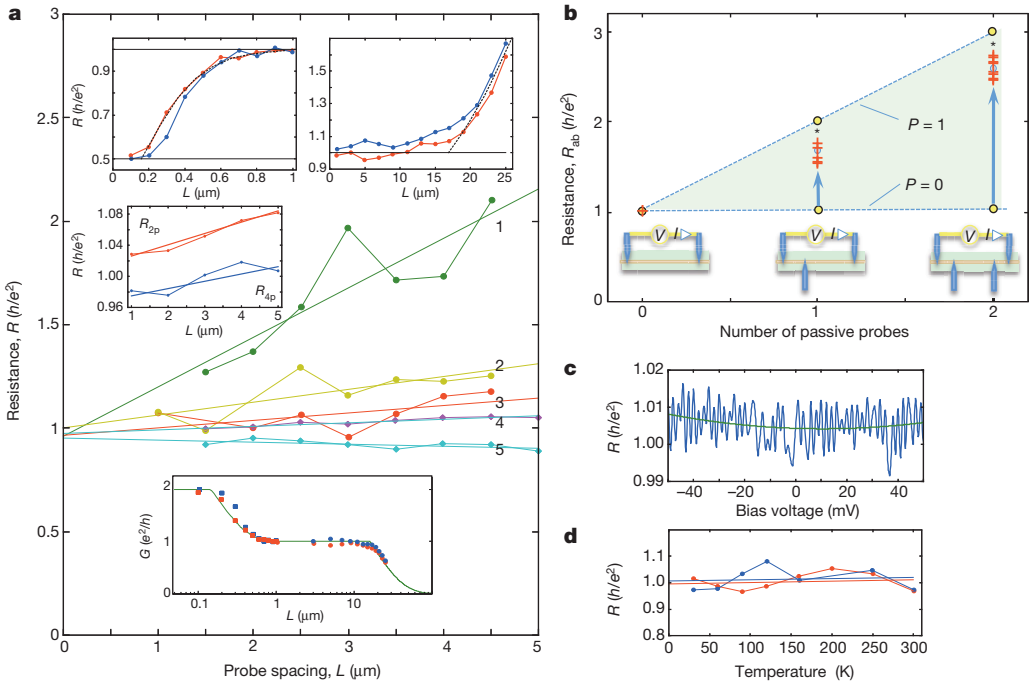


Figure 3 | Multiprobe *in situ* transport measurements of sidewall ribbons. **a**, Resistance versus probe spacing L. Linear fits extrapolate to $R_0 = h/e^2$. Slopes correspond to $R' = 6.2$, 1.6, 0.92, 0.44 kΩ μm^{-1}, corresponding to the mean free path $\lambda_0 = 4.2$, 28, 16, 58 μm (following transmission $\mathrm{Tr} \approx (1 + L/\lambda_0)^{-1}$). Line 5 is consistent with zero slope. Numbered traces are as follows: 1, UHV annealed at 1,100 °C for 15 min; 2, UHV annealed at 1,150 °C for 15 min; 3–5, re-annealed at 1,150 °C for 15 min. Middle inset, comparison of two-probe (2p) and four-probe (4p) measurements. Upper insets, nonlinear resistance increases observed at $L = 160$ nm and at $L = 16$ μm in two different ribbons measured at room temperature, presented as $G(L)$ in the lower inset. Fit is explained in main text. **b**, Effect of passive probes contacting sidewall ribbons. The resistance essentially doubles with one passive probe and triples with two passive probes. This property of ballistic conductors explains why four-probe and two-probe measurements yield essentially identical resistance values for ballistic wires. Ideal invasive probe ($P = 1$) and non-invasive probe ($P = 0$) limits are indicated. Theoretical[18,19] values (for RR = 0.95) indicated by asterisks. **c**, Resistance R_{4p} of a typical ribbon for $L = 5$ μm versus bias voltage V_b. **d**, Resistance versus temperature for the same ribbon, showing less than 10% variation from 30 K to 300 K.

$R(L)$ was measured at 300 K from 1 μm to the apparatus limit, $L = 25$ μm (Fig. 3a upper right inset). A second nonlinear increase is observed for $L > L^*_{0+} = 16$ μm. These nonlinearities are consistent with an exponential conductance decrease given by $G = G_0\exp(1 - L/L^*_{0\pm})$ for $L \geq L^*_{0\pm}$ as shown in Fig. 3a (dashed line in upper right inset). It is intriguing that $E^*_{01}/k_B = \pi\hbar c^*/L^*_{0-} = 150$ K, suggesting that longitudinal excitations may be involved (equation (1)).

We next discuss the transport properties of a fixed-geometry sample (such as shown in Fig. 1c)[7]. Further examples are given in Supplementary Information. Sample A is a top-gated 39-nm-wide 1.6-μm-long graphene sidewall ribbon. The ribbon is seamlessly connected to micrometre-scale graphene pads to the left and right. Each pad is bonded to two wires, facilitating four-point transport measurements. Resistances around 20 kΩ are measured with better than 0.1 Ω precision (corresponding to $\delta G < 5 \times 10^{-6}\ G_0$) using standard low-frequency lock-in techniques (13 Hz, 100 nA $< I <$ 1 μA)[7]. Temperatures are measured with 2 mK precision. The charge density $n(V_g)$ is adjusted by applying a gate voltage, V_g: we find that $n(V_g) = -0.95 \times 10^{12}\ V_g\ \mathrm{cm}^{-2}\ \mathrm{V}^{-1}$, as determined from a Hall bar on the same substrate[7], so that $E_F = -0.11\ V_g^{1/2}\ \mathrm{eV\ V}^{-1/2}$.

Figure 4c shows the conductance $G(V_g)$ as function of gate voltage for several temperatures, and can be globally explained in the Landauer picture with $\mathrm{Tr}_n(E_F) = \mathrm{Tr}_n\theta(|E_F| - |E_{n0}|)$ (θ is the step function)[10]. The minimum conductance $G = 0.95G_0$ at $V_g \approx 0$ and $T = 4.2$ K is consistent with a single ballistic channel in charge-neutral graphene with $\lambda_{0+} = 22$ μm (from $\mathrm{Tr}_n \approx (1 + L/\lambda_n)^{-1}$). The conductance increase with increasing V_g corresponds to the opening of the $|n| \geq 1$ subbands[20], as diagrammatically shown in Fig. 4b. From the $G(V_g)$ slope we deduce that $T_{|n|\geq 1} = 0.035$, so that $\lambda_{|n|\geq 1} = 60$ nm. We note that each of the curves can be displaced vertically to overlap the others. This is consistent with the Landauer picture and Fig. 3, if we assume that $T_{|n|\geq 1}$ are temperature independent and that only $T_{n=0-}$ is dependent on temperature, as indicated in Fig. 4b. (This subband-dependent temperature effect is also seen in exfoliated graphene ribbons, see Supplementary Fig. 2.)

Note that the large asymmetry with respect to V_g (Fig. 4a) is caused by the np/pn junctions (see, for example, ref. 21). For $V_g < 0$, the ribbon is p doped while the leads are slightly n doped[7,16]. Because $\lambda_{|n|\geq 1} \approx 50$ nm is larger than the junction width, the junctions represent a significant barrier for the gapped $n \neq 0$ subbands; however, the ungapped $n = 0$ subbands are not affected[21].

The conductance $G(T)$ increases monotonically with increasing temperature, as shown in Fig. 4e. In these experiments samples were cooled from 120 K to 4 K over 10 h, and two measurements were performed per second. The conductance is described to remarkable precision by

$$G(T) = \alpha\frac{e^2}{h}\left[1 + 0.5\exp\left\{-\left(\frac{T^*}{T_{el} - T_0}\right)^{1/2}\right\}\right] \qquad (2)$$

where T_{el} is the effective electronic temperature (in this case, it is equal to the sample temperature), $\alpha = 0.922$, $T^* = 21.5$ K and $T_0 = 2.2$ K. The difference $\delta G(T)$ between the fit and data is within $0.0015G_0$ (0.1%) from T = 4 K to T = 120 K (Fig. 4e lower inset). Remarkably, the activation temperature T* is related to the sample length L, with $T^* = 1.4\pi\hbar c^*/k_B L = 20.9$ K for $L = 1.6$ μm. Equation (2) applies to samples B, C and D (see Supplementary Information) as well. Specifically, for sample B (Supplementary Fig. 4), $\alpha = 0.31$ and the measured T^* is $T^*_m = 29$ K; for $L = 1.06$ μm, the calculated T^* is $T^*_p = 1.4\pi\hbar c^*/k_B L = 31$ K; for sample D (Supplementary Fig. 5), $\alpha = 0.63$ and $T^*_m = 87$ K; for $L = 0.36$ μm, $T^*_p = 93$ K. For the ring structure (Fig. 1d), $T^*_m = 6 \pm 1$ K, and the measured contact-to-contact distance is 5 μm (following half a turn

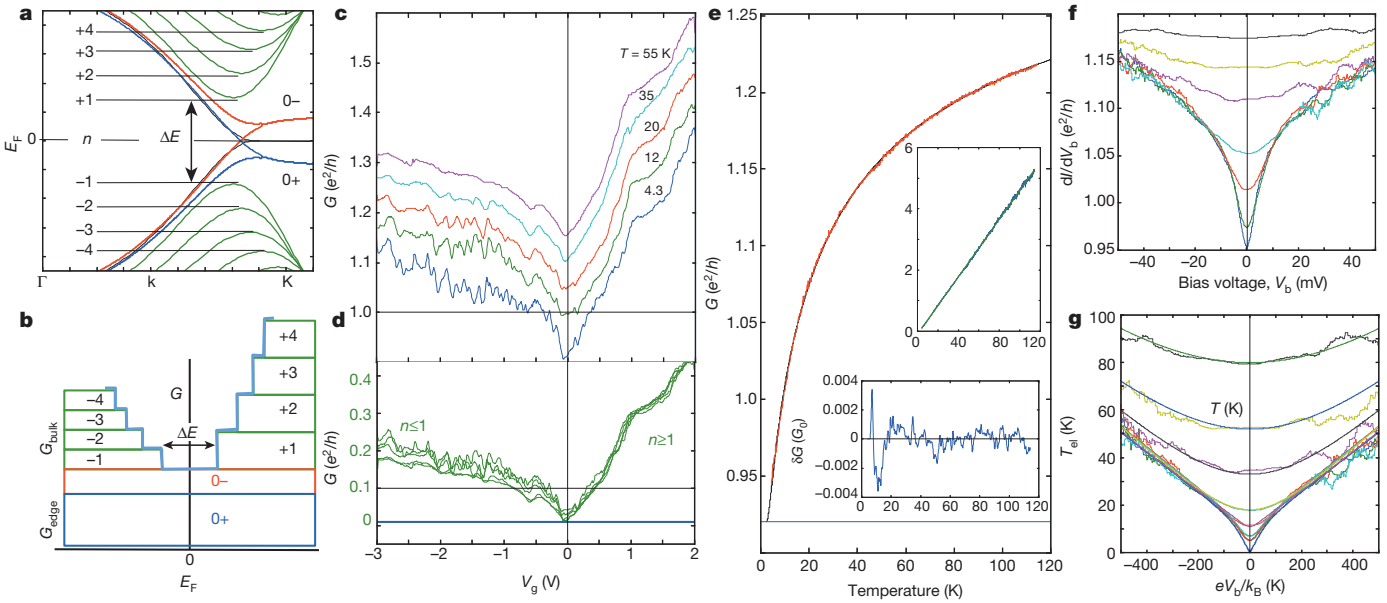


Figure 4 | Gated ribbon transport (see Fig. 1c). **a**, Schematic band structure. **b**, Conductance increase with increasing V_g is due to opening of subbands; for $|E_F| \leq E_1$, only the $n = 0\pm$ subbands contribute; the conductance increase with increasing temperature is due to the $0-$ subband. **c**, Conductances $G_T(V_g)$ for various temperatures. Minimum conductance at $V_g = 0$ corresponds to charge-neutral ribbons ($E_F = 0$). **d**, Decomposition of $G_T(V_g) = G_{0+} + G_{0-}(T) + G_n(V_g)$. The $|n| \geq 1$ bands ($E_F \neq 0$) show no temperature dependence (apart from weak oscillations), as seen by the collapse of all the curves onto a single curve. **e**, Four-point G versus T (red curve) superimposed on the theoretical curve from equation (2) (blue curve), showing an excellent fit. Upper inset, $[\ln(2G/\alpha G_0 - 2)]^{-2}$ versus T; the inverse slope of the resulting line is $T^* = 21.5$ K and the zero intercept is $T_0 = 2.2$ K (equation (2)). Lower inset, difference $\delta G(T)$ between equation (2) and experiment. **f**, dI/dV_b versus bias voltage V_b; from bottom to top, T (in K) is 4.2, 7, 12, 20, 35, 55, 80. **g**, Same data as in **f** plotted as a function of T_{el}; $T_{el}(G)$ is determined from data in **e**, and $T_{Vb} = eV_b/k_B$. Fits correspond to equation (3). For $T_{el} < 15$ K, $v = 5$; for larger T_{el}, $v = 10$.

of the circle), from which $T_p^* = 6.7$ K. This establishes the inverse L dependence of T^* in these samples (α and T^* appear to be unrelated), and implies that T^* is related to E_{01} (see equations (1) and (2)).

The increase of conductance with increasing bias voltage V_b (Fig. 4f) is attributed to electronic heating[3]. Using the conductance (equation (2)) as a thermometer allows us to determine $T_{el}(T, V_b)$ from the conductance. Moreover, we define $T_{Vb} = eV_b/k_B$, to enable us to plot T_{el} versus T_{Vb}, as shown in Fig. 4g. The data are fitted with:

$$T_{el} = \sqrt{T^2 + (T_{Vb}/v)^2} \qquad (3)$$

Figure 4g shows a good fit for all temperatures with the coefficient $v = 5$ for $T_{el} < 15$ K and $v = 12$ for $T_{el} > 15$ K. This behaviour is observed in samples B, C and D, and also in carbon nanotubes (Supplementary Figs 3, 4, 5, 6, 8).

Equation (2) resembles Mott's expression for one-dimensional variable range hopping[22]. It is therefore reasonable to expect a related mechanism. Following Mott's heuristic argument, charge carriers in disordered conductors ballistically hop from one scattering centre to the next, separated by a hopping distance L^*, following the path of least resistance, that is, one with the largest transmission coefficient. This transmission coefficient involves the product Z of two competing terms. One is the Boltzmann factor of the lowest longitudinal mode between the hopping endpoints, with activation energy $E_{01}^* = \pi\hbar c^*/L^*$, and the other is determined by the lifetime of the charge carrier τ. Consequently $Z = \exp(-E_{01}^*/k_BT)\exp(-t/\tau) = \exp(-\pi\hbar c^*/L^*k_BT)\exp(-L^*/c^*\tau)$, with t the carrier transit time over the distance L^*. We note that Z is a maximum for $L^* = \sqrt{(\pi\hbar c^{*2}\tau/k_BT)}$, so that $Z = \exp -\sqrt{(T^*/T)}$, where $T^* = 4\pi\hbar/k_B\tau$. From equation (2) and the inverse relation between T^* and L, we conclude that $\tau = 2.8L/c^*$, that is, of the order of (but larger than) the ballistic transit time through the ribbon. This rough estimate suggests that transport involves thermally activated longitudinal modes of the graphene ribbon. The dependence on bias voltage indicates that impinging electrons produce hot charge carriers that overcome the activation barrier. The factor v relates to the efficiency of this heating process.

The transport properties reported in Fig. 3 agree with those reported in Fig. 4, both showing two-component behaviour, each component contributing at most $1G_0$ to the conductance. The mean free path of the first component typically exceeds tens of micrometres and its resistance increases nonlinearly above 16 μm. The second component is unusual. Transport is ballistic between scattering centres on the length scale L^*, which can be larger than the ribbon length L. Transport appears to be thermally activated with an activation barrier determined by the lowest longitudinal mode between the scattering centres (that is, E_{01}^*). This channel exhibits a positive linear magnetoconductance (discussed in Supplementary Information) that typically saturates at $B = 2$ T.

Room-temperature single-channel ballistic conduction in the $n = 0+$ mode was first observed two years ago in epitaxial graphene sidewall ribbons[7]. Single-channel ballistic transport in the $n = 0+$ channel is supported by the robust value of the quantum 'contact resistance' (at $L = 0$); the independence of the resistance with respect to length, temperature and bias voltage; the equivalence of two- and four- probe resistance measurements; and resistance doubling and tripling due to passive probes. Its insensitivity to gate voltages implies that this component is derived from a metallic subband[13]. Conduction in the $n = 0-$ mode appears to involve thermally activated transport (equation (2)) that is ballistic between widely spaced scattering centres. The separation of scattering centres can exceed the ribbon length in fixed geometry samples, where the ribbon is seamlessly connected to graphene leads. The activation barrier is found to be inversely proportional to the ribbon length. In addition, an energy gap equivalent to $T_0 = 2.2$ K appears to be involved. The conductance bias voltage dependence is explained in terms of electronic heating. Its insensitivity to gate voltages implies that it too is derived from a metallic subband. The $n = 0-$ mode also exhibits linear positive magnetoconductance. If the conductance increase is

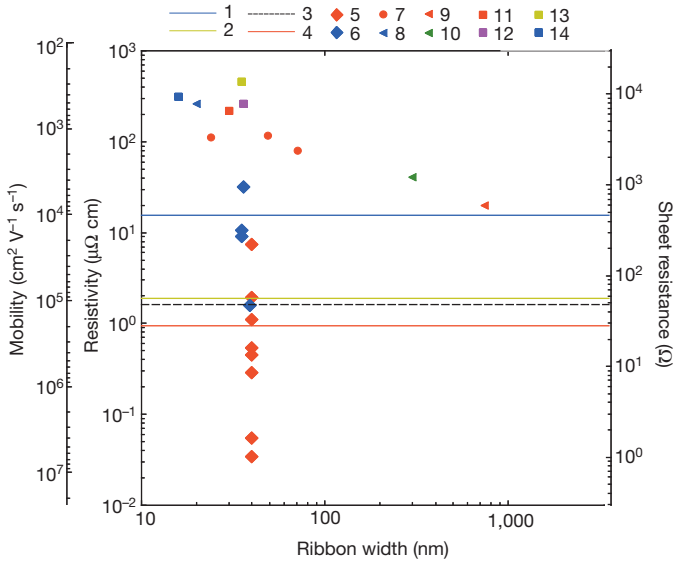


Figure 5 | Comparison with other work. Transport properties of epitaxial graphene nanoribbons reported here (5 and 6) are compared with exfoliated graphene as follows: 1, two-dimensional graphene on SiO_2 (ref. 25); 2, ultrahigh-mobility graphene on BN (ref. 26); 3, bulk silver; 4, the theoretical ideal graphene limit[25], and lithographically prepared graphene nanoribbons at a charge density $n_s = 10^{12}\,cm^{-2}$ at $T = 30$ K; 7, ref. 2; 8, ref. 4; 9, ref. 27; 10, ref. 28; 11, ref. 29; 12, ref. 3; 13, ref. 5; 14, ref. 30. Sheet resistances (R_{square}) for graphene ribbons are determined by multiplying reported resistances (in R_0 units) by W/L. For back-gated graphene, the resistivity at $V_g = V_D \pm 14$ V (corresponding to $n_s = \pm 10^{12}\,cm^{-2}$) is reported, where V_D locates the Dirac point. Resistivities ρ correspond to $\rho = dR_{square}$, where $d = 3 \times 10^{-8}$ cm corresponds to a monolayer thickness. Mobilities μ are determined from the definition $\mu = (n_s e R_{square})^{-1}$ at a charge density of $10^{12}\,cm^{-2}$. For two-dimensional and bulk materials, bulk values are used (not adjusted for finite-size effects).

attributed to a rise of the chemical potential due to the magnetic field, then this implies that the charge carriers have magnetic moments of about $5\mu_B$ (see Supplementary Information). At room temperature, both channels apparently exhibit exponential resistance increases when their lengths increase beyond a threshold (16 μm for the 0+ mode and 160 nm for the 0− mode).

The properties of the ballistic modes are distinct from those of the other subbands ($|n| \geq 1$). Although the mean free paths of the $|n| \geq 1$ subbands (60 nm) are much larger than those observed in exfoliated graphene ribbons (~5 nm, see Supplementary Fig. 2), they are in line with those observed by others in two-dimensional graphene samples (Fig. 5). Clearly scattering mechanisms that apply to normal Dirac electrons do not apply to the ballistic modes discussed here. Because both modes have a conductance of $1G_0$ they probably represent distinct longitudinal modes of the $n = 0\pm$ subbands (Fig. 3a), but they cannot both be derived from the same one-dimensional subband. Apparently the $n = 0+$ mode represents the ground-state longitudinal mode derived from the $n = 0+$ subband and the $n = 0-$ mode represents an excited state derived from the $n = 0-$ subband. The fact that properties of these two modes are so different is not expected and points to a broken symmetry. We emphasize that very similar properties are observed in carbon nanotubes[23] (Supplementary Information).

We have presented compelling evidence for single-channel ballistic transport in long epitaxial graphene nanoribbons. This implies that both spin and valley degeneracy is lifted. Moreover, exceptional transport is observed only in the $n = 0$ subband of the graphene ribbon and it is distinct from the behaviour of the $|n| \geq 1$ subbands that have much shorter mean free paths. We note that for exfoliated, lithographically patterned ribbons, the inverse is true: the $n = 0$ subbands have much shorter mean free paths than the $|n| \geq 1$ subbands, resulting in a mobility gap. Consequently, the properties we observe are not simply the result of clean samples but possibly reflect correlation effects[24] that may be particularly important in the $n = 0$ subbands. Regardless of its origin, it likely that room-temperature ballistic transport will play an important part in future graphene nanoelectronics.

METHODS SUMMARY

For ribbons in Figs 2 and 3, trenches 20 nm deep were etched along the $[1\bar{1}00]$ direction. Samples in Fig. 2 were annealed at 1,600 °C for 15 min (ref. 7). Samples in Fig. 3 were heated at 1,300 °C in Ar (4×10^{-5} mbar), and then in UHV for 15 min at 1,100 °C (green dots, Fig. 3a) or at 1,150 °C (all others—see Supplementary Information). The natural-step ribbon[7] (Fig. 1c and Fig. 4) was connected at each end to two 200-nm-deep graphitized trenches 1 μm apart that serve as current leads and voltage probes.

All ribbons were characterized with AFM, conducting AFM, SEM and electrostatic force microscopy, to determine ribbon widths and sidewall slopes (see ref. 7 for details, and also refs 6,15). The Soleil synchrotron facility (Cassiopée ARPES beam line) was used for graphene band structure measurements (Supplementary Fig. 1), to determine the number of graphene layers and the sidewall slopes[7,16].

Low-temperature ($T = 77$ K) STM-STS measurements (Fig. 2) were performed (in Nancy) in a UHV chamber coupled to a preparation chamber. For STM, bias voltages are relative to the grounded tip. STS spectra were acquired with a PtIr tip ($V_{bias} = 70$ mV, 1,100 Hz) and a lock-in current detection, in open feedback loop conditions.

The multi-probe measurements in Fig. 3 were performed in an Omicron Nanoprobe UHV system (in Hannover), from 30 K to 300 K, using W tips positioned using a built-in SEM. The STS set point (Fig. 1d) was 2 V/0.1 nA. The resistance between neighbouring ribbons was >450 kΩ (Supplementary Fig. 11).

Results in Fig. 4 were obtained (in Atlanta) in a Janis variable temperature cryostat (4–300 K) with a 9-T magnet, using standard low-frequency lock-in techniques (13 Hz, 100 nA < I < 1 μA)[7]. For $G(T)$, samples were cooled from 120 K to 4 K over 10 h, with a rate of two measurements per second.

Received 28 August; accepted 11 December 2013.
Published online 5 February 2014.

1. Berger, C. *et al.* Ultrathin epitaxial graphite: 2D electron gas properties and a route toward graphene-based nanoelectronics. *J. Phys. Chem. B* **108,** 19912–19916 (2004).
2. Han, M. Y. Özyilmaz, B. Zhang, Y. & Kim, P. Energy band-gap engineering of graphene nanoribbons. *Phys. Rev. Lett.* **98,** 206805 (2007).
3. Han, M. Y., Brant, J. C. & Kim, P. Electron transport in disordered graphene nanoribbons. *Phys. Rev. Lett.* **104,** 056801 (2010).
4. Chen, Z. H., Lin, Y. M., Rooks, M. J. & Avouris, P. Graphene nano-ribbon electronics. *Physica E* **40,** 228–232 (2007).
5. Todd, K., Chou, H. T., Amasha, S. & Goldhaber-Gordon, D. Quantum dot behavior in graphene nanoconstrictions. *Nano Lett.* **9,** 416–421 (2009).
6. Sprinkle, M. *et al.* Scalable templated growth of graphene nanoribbons on SiC. *Nature Nanotechnol.* **5,** 727–731 (2010).
7. Ruan, M. *Structured Epitaxial Graphene for Electronics.* PhD thesis, Georgia Inst. Technol. (2012); available at http://hdl.handle.net/1853/45596.
8. Castro Neto, A. H., Guinea, F., Peres, N. M. R., Novoselov, K. S. & Geim, A. K. The electronic properties of graphene. *Rev. Mod. Phys.* **81,** 109–162 (2009).
9. Berger, C. *et al.* Electronic confinement and coherence in patterned epitaxial graphene. *Science* **312,** 1191–1196 (2006).
10. Datta, S. *Electronic Transport in Mesoscopic Systems* (Cambridge Univ. Press, 1995).
11. Nakada, K., Fujita, M., Dresselhaus, G. & Dresselhaus, M. S. Edge state in graphene ribbons: nanometer size effect and edge shape dependence. *Phys. Rev. B* **54,** 17954–17961 (1996).
12. Wakabayashi, K., Takane, Y. & Sigrist, M. Perfectly conducting channel and universality crossover in disordered graphene nanoribbons. *Phys. Rev. Lett.* **99,** 036601 (2007).
13. Frank, S., Poncharal, P., Wang, Z. L. & de Heer, W. A. Carbon nanotube quantum resistors. *Science* **280,** 1744–1746 (1998).
14. de Heer, W. A. *et al.* Large area and structured epitaxial graphene produced by confinement controlled sublimation of silicon carbide. *Proc. Natl Acad. Sci.* **108,** 16900–16905 (2011).
15. Hu, Y. *et al.* Structured epitaxial graphene: growth and properties. *J. Phys D* **45,** 154010 (2012).
16. Hicks, J. *et al.* A wide-bandgap metal–semiconductor–metal nanostructure made entirely from graphene. *Nature Phys.* **9,** 49–54 (2013).
17. Norimatsu, W. & Kusunoki, M. Formation process of graphene on SiC (0001). *Physica E* **42,** 691–694 (2010).
18. Büttiker, M. Four terminal phase coherent conductance. *Phys. Rev. Lett.* **57,** 1761–1764 (1986).
19. de Picciotto, R., Stormer, H. L., Pfeiffer, L. N., Baldwin, K. W. & West, K. W. Four-terminal resistance of a ballistic quantum wire. *Nature* **411,** 51–54 (2001).

20. Tombros, N. *et al.* Quantized conductance of a suspended graphene nanoconstriction. *Nature Phys.* **7,** 697–700 (2011).
21. Huard, B. *et al.* Transport measurements across a tunable potential barrier in graphene. *Phys. Rev. Lett.* **98,** 236803 (2007).
22. Mott, N. F. Conduction in non-crystalline materials. III. Localized states in a pseudogap and near extremities of conduction and valence bands. *Phil. Mag.* **19,** 835–852 (1969).
23. Schonenberger, C., Bachtold, A., Strunk, C., Salvetat J. P. & Forro, L. Interference and interaction in multi-wall carbon nanotubes. *Appl. Phys. A* **69,** 283–295 (1999).
24. Das Sarma, S., Adam, S., Hwang, E. H. & Rossi, E. Electronic transport in two-dimensional graphene. *Rev. Mod. Phys.* **83,** 407–470 (2011).
25. Chen, J. H., Jang, C., Xiao, S. D., Ishigami, M. & Fuhrer, M. S. Intrinsic and extrinsic performance limits of graphene devices on SiO_2. *Nature Nanotechnol.* **3,** 206–209 (2008).
26. Mayorov, A. S. *et al.* Micrometer-scale ballistic transport in encapsulated graphene at room temperature. *Nano Lett.* **11,** 2396–2399 (2011).
27. Huard, B., Stander, N., Sulpizio, J. A. & Goldhaber-Gordon, D. Evidence of the role of contacts on the observed electron-hole asymmetry in graphene. *Phys. Rev. B* **78,** 121402R (2008).
28. Lemme, M., Echtermeyer, T. J., Baus, M. & Kurz, H. A graphene field effect device. *IEEE Electron Device Lett.* **28,** 282–284 (2007).
29. Lin, Y. M., Perebeinos, V., Chen, Z. H. & Avouris, P. Electrical observation of subband formation in graphene nanoribbons. *Phys Rev B* **78,** 161409(R) (2008).
30. Wang, X. R. *et al.* Graphene nanoribbons with smooth edges behave as quantum wires. *Nature Nanotechnol.* **6,** 563–567 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements C.T. thanks the German Research Foundation Priority Program 1459 'Graphene' for financial support. C.B., E.H.C. and W.A.d.H. thank R. Dong, P. Goldbart, Z. Guo, J. Hankinson, J. Hicks, Y. Hu, J. Kunc, M. Kindermann, D. Mayou, M. Nevius, J. Palmer, A. Sidorov and P. de Heer for assistance and comments. C.B., E.H.C. and W.A.d.H. thank the AFOSR, NSF (MRSEC – DMR 0820382), W. M. Keck Foundation and Partner University Fund for financial support. Work at ORNL was supported by the Scientific User Facilities Division, BES of the DOE.

Author Contributions J.B. and F.E. produced samples and performed the *in situ* transport experiments in Hannover relating to Fig. 3. C.T. performed and supervised the transport experiments in Fig. 3, discussed the data and commented on the paper. M.R. produced the samples and performed transport experiments shown in Fig. 4 and Supplementary Figs 3–6. E.H.C., A.T. and A.T.-I. performed ARPES experiments, and A.T. and M.S. the STM and STS experiments. Z.J. performed confirming spin transport measurements and contributed to C-AFM results shown in Supplementary Fig. 6b. A.-P.L. performed earlier SPM measurements. W.A.d.H. conceived and supervised the experiment and interpreted the data. C.B. supervised and performed the Atlanta based experiments. W.A.d.H. and C.B. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to W.A.d.H. (walt.deheer@physics.gatech.edu).

LETTER

doi:10.1038/nature12992

Prodigious degassing of a billion years of accumulated radiogenic helium at Yellowstone

J. B. Lowenstern[1], W. C. Evans[1], D. Bergfeld[1] & A. G. Hunt[2]

Helium is used as a critical tracer throughout the Earth sciences, where its relatively simple isotopic systematics is used to trace degassing from the mantle, to date groundwater and to time the rise of continents[1]. The hydrothermal system at Yellowstone National Park is famous for its high helium-3/helium-4 isotope ratio, commonly cited as evidence for a deep mantle source for the Yellowstone hotspot[2]. However, much of the helium emitted from this region is actually radiogenic helium-4 produced within the crust by α-decay of uranium and thorium. Here we show, by combining gas emission rates with chemistry and isotopic analyses, that crustal helium-4 emission rates from Yellowstone exceed (by orders of magnitude) any conceivable rate of generation within the crust. It seems that helium has accumulated for (at least) many hundreds of millions of years in Archaean (more than 2.5 billion years old) cratonic rocks beneath Yellowstone, only to be liberated over the past two million years by intense crustal metamorphism induced by the Yellowstone hotspot. Our results demonstrate the extremes in variability of crustal helium efflux on geologic timescales and imply crustal-scale open-system behaviour of helium in tectonically and magmatically active regions.

Despite its notable mobility, the light element He can be stored over geologically extensive periods[3]. It is found in considerable quantities in some natural-gas reservoirs, where it remains stored for millions of years[3,4]. Yet it can rapidly traverse the crust in tectonically active regions where mantle ^{3}He leaks to the surface[5]. Recognition of He accumulations within sedimentary basins and oil fields spurred considerable research to assess whether the He migrates during steady-state degassing of the entire crust, or during more transient dynamic episodes associated with increased heat flow or tectonism[3,6,7]. Tracking the magnitude of accumulation of He within crustal rocks themselves, and the rapidity with which He can be purged, has not yet been fully explored.

Studies of He flux from active tectonic and magmatic regions usually focus on the mantle-derived component; those that include crustal flux have typically explored point sources limited in extent[7,8]. The Yellowstone plateau volcanic field is often viewed as a mantle hotspot with well-known high ^{3}He/^{4}He ratio (R) in gas from fumarolic vents and hot-spring waters[2,9]. Decades ago, workers recognized that Yellowstone's mantle He signature is diluted considerably by crustal ^{4}He, but at that time the mass flux of ^{4}He implied by this dilution was not appreciated.

Over the past ten years, we have combined studies of gas and isotope chemistry with measurements of CO_2 emission rates from thermal areas across Yellowstone National Park[10–13]. One clear result has been that the very high CO_2 emissions require considerable efflux of He. Gas emissions are concentrated in and around the Yellowstone caldera, which was formed during the last major caldera-forming eruption, 640 kyr ago. The caldera is the geographic focus for continuing uplift–subsidence cycles[14], presumably because its location coincides with the present-day input ($>0.1\,km^3\,yr^{-1}$) of mantle basalt into the crust[15,16]. The inferred mantle input is consistent with ^{3}He/^{4}He ratios (Fig. 1) that reach values more than sixteen times the atmospheric ratio (that is, $16R_a$) within the caldera. However, much lower values prevail in some parts of the caldera and the surrounding volcanic field.

Notably, ^{3}He/^{4}He ratios in gas from fumaroles and bubbling pools decrease as He concentrations increase (Fig. 2a). This is consistent with dilution of the mantle He signature by radiogenic crustal ^{4}He. Reduced R/R_a and increased He concentrations also correlate with increasing CH_4 and decreasing $\delta^{13}C$ CH_4 (Fig. 2), which are clear signs of crustal input[8]. ($\delta^{13}C$ (‰) $= [(^{13}C/^{12}C)_{sample}/(^{13}C/^{12}C)_{standard} - 1] \times 1{,}000$, where the standard is PeeDee Belemnite. The reference to CH_4 reflects that we are measuring the carbon isotopic ratio in CH_4.)

All these trends reflect in part the influence of Eocene Absaroka Supergroup rocks, which are present in the eastern part of the park and contribute gas with high ethane/methane ratios, low $\delta^{13}C$ CH_4 values and greater methane concentrations than are found in the rest of Yellowstone[10,17]. But the trends hold for all of Yellowstone up to the highest measured R/R_a values, supporting the interpretation that crustal ^{4}He input is the main reason all R/R_a values fall below the value of 22 inferred for the Yellowstone hotspot endmember[18].

At Yellowstone, the highest He concentrations yet found come from the Heart Lake Geyser Basin[12], at the south boundary of the caldera, adjacent to the active East Mount Sheridan faults that expose Palaeozoic metasediments. Reported He/CO_2, R/R_a, and CO_2 emissions[12] (Extended Data Figs 1 and 2) permit estimation of the crustal He flux from the geyser basin:

$$q^{C}_{He} = 3.65 \times 10^{8} \frac{Q_{CO_2}}{M_{CO_2}} \phi \frac{X_{He}}{X_{CO_2}}$$

where Q_{CO_2} is the measured emission of CO_2 in tonnes per day, M_{CO_2} is the molecular weight of CO_2, X_{He}/X_{CO_2} is the molar ratio and ϕ is the fraction of crustal He in the sample (Methods). The expression yields q^{C}_{He} in moles per year.

Calculated total He emissions from the Heart Lake Geyser Basin are $6{,}900\,mol\,yr^{-1}$ (Table 1). Given the very low R/R_a of gas samples from the Heart Lake area (1.1 to 2.9), and the presumed high R/R_a of the hotspot endmember (22), q^{C}_{He} is equal to $5{,}970\,mol\,yr^{-1}$ of crustal ^{4}He.

The rate of ^{4}He production due to α-decay of U and Th in the entire crust beneath Yellowstone (Q^{C}_{He} in moles per year) is equivalent to

$$Q^{C}_{He} = \frac{M_Y}{N_A}\alpha$$

where M_Y is the mass of the crust beneath Yellowstone in grams, N_A is Avogadro's constant and α is the crustal production of ^{4}He in atoms per gram per year. In turn

$$\alpha = 3.115 \times 10^{6}[U] + 1.272 \times 10^{5}[U] + 7.710 \times 10^{5}[Th]$$

where [U] and [Th] are respectively the concentrations of U (^{235}U and ^{238}U, hence the two terms) and Th in the crust in parts per million by weight[19]. Assuming an average crustal composition[20] and a 43-km-thick crust (Methods), the production rate of ^{4}He beneath Yellowstone is $17{,}100\,mol\,yr^{-1}$. The crustal ^{4}He flux from the 0.8-km^2 Heart Lake Geyser Basin is thus equivalent to ~35% of this whole-crust (WC) production rate beneath the 9,000-km^2 park.

[1]US Geological Survey, Menlo Park, California 94025, USA. [2]US Geological Survey, Denver, Colorado 80225, USA.

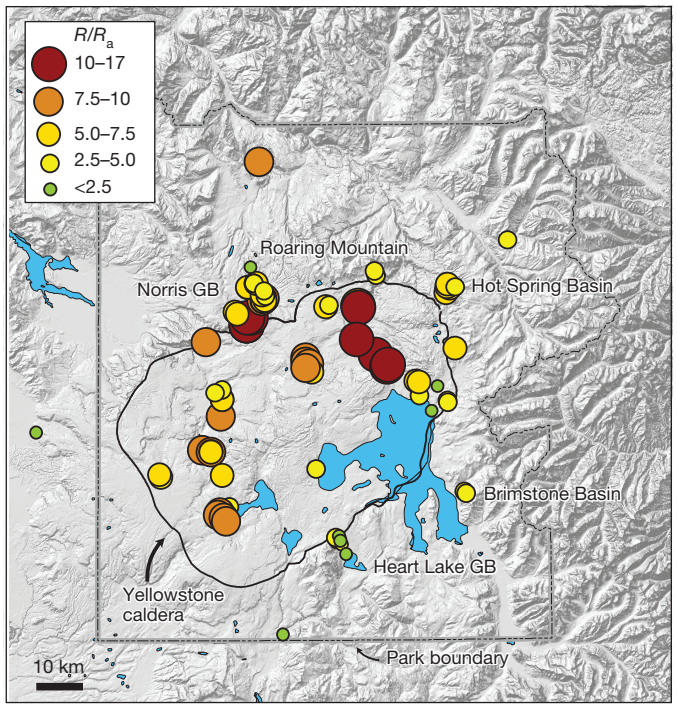


Figure 1 | Map of Yellowstone. Circles denote sample locations colour-coded and sized by He isotope composition (R/R_a) as depicted in the key (all values were corrected for minor air contamination). Thermal areas discussed in the text are shown; the relevant samples are immediately to the left of the area name, except for Norris Geyser Basin (to right). Data are from refs 10, 16.

Other degassing areas around the periphery of the caldera also have very low R/R_a owing to evident addition of crustal He (refs 9, 10 and 18). Combining annual estimated emissions for Hot Spring Basin[11], Brimstone Basin[13] and Roaring Mountain[16] with Heart Lake Geyser Basin yields the equivalent of 28WC (Table 1). Yet these four locations, which have a combined area of ~5 km^2, represent <10% of the total heat[21] and gas[16] flux at Yellowstone. We also estimate the total radiogenic He flux from the entire Yellowstone hydrothermal system. The median air-corrected R/R_a of gas from 83 localities within and outside the caldera is 6.9 (Extended Data Fig. 1), which is far less than the value for the hotspot endmember[18], of 22. Our best park-wide estimate of crustal ^{4}He flux is 597WC, or 21 times greater than the four thermal areas estimated above.

It is possible that some of the trend in Fig. 2a could be due to a second mantle endmember such as very old subcontinental lithospheric mantle with $^3He/^4He = 4R_a–6R_a$, which is thought by many to be an important magma source in Snake River Plain magmatism[22]. Involvement of subcontinental lithospheric mantle would not explain the observed trends in CH_4/CO_2 or $\delta^{13}C$ CH_4 versus R/R_a (Fig. 2), nor the higher NH_3 concentration and C_2H_6/CH_4 in gases with low R/R_a (ref. 10). Methane and C_2H_6 are normally produced at temperatures less than 300 °C and by thermogenesis from organic sediments, and are unstable in high-temperature, mantle-derived volcanic gas[23]. Nevertheless, we calculate crustal ^{4}He fluxes assuming both the Yellowstone hotspot endmember, with $R = 22R_a$ (ref. 18), as well as a bulk mixed mantle with $R = 11R_a$ (Table 1). This value represents the lowest $^3He/^4He$ ratio found in a suite of basalt-hosted olivines from nine localities throughout the eastern Snake River Plain[22]. Using both mantle endmembers, we constrain the total crustal ^{4}He emissions from Yellowstone at between 86WC and 1,970WC, with a best estimate of 597WC. If the whole crust beneath Yellowstone were particularly U- and Th-rich (4 parts U per million and 12 parts Th per million), this would decrease to 238WC. Degassing of the Yellowstone hydrothermal system thus yields tens to hundreds of times more radiogenic He than can be supported by the underlying crust. If similar He emissions were discharged over the 2-Myr history of the Yellowstone plateau volcanic field, then between 172 and 3,940 Myr of accumulated radiogenic He must have been emitted during that time (best estimate of 1,194 Myr).

Such great volumes of He would be particularly difficult to explain if the crust were composed entirely of young volcanic rock. Though Pleistocene volcanic rocks crop out at the surface in the caldera, and Eocene rocks of the Absaroka Supergroup dominate at the park's

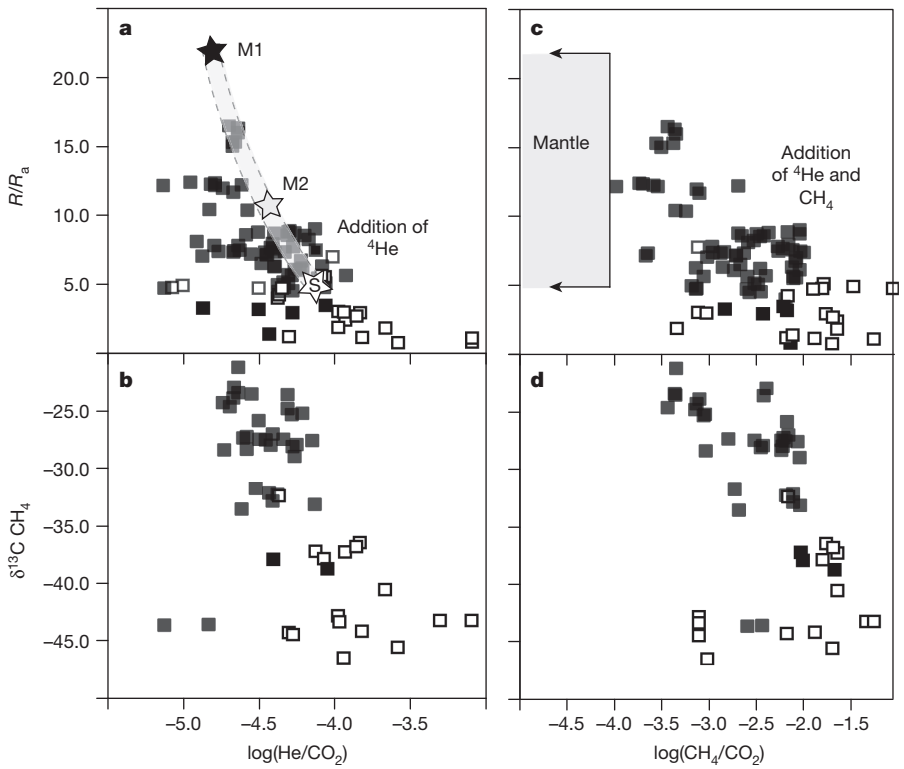


Figure 2 | Concentration and isotope ratios for Yellowstone gases. **a**, $^3He/^4He$ (as R/R_a) versus $\log(He/CO_2)$; **b**, $\delta^{13}C$ CH_4 versus $\log(He/CO_2)$; **c**, $^3He/^4He$ (as R/R_a) versus $\log(CH_4/CO_2)$; **d**, $\delta^{13}C$ CH_4 versus $\log(CH_4/CO_2)$. CO_2 makes up >90% of virtually all Yellowstone gas samples (steam excluded). Open squares show the extracaldera thermal areas denoted in Fig. 1 plus those clearly emerging through Absaroka volcaniclastic rocks at the caldera's eastern edge. Filled squares are intracaldera thermal features plus those of the adjacent Norris Geyser Basin. In **a**, mantle endmembers M1 (ref. 18) and S (ref. 22) (subcontinental lithospheric mantle) are mixed along the grey bar to create M2, which is used along with M1 in flux calculations (Methods). Locations of M1, M2 and S in **b**–**d** are unknown, although $\log(CH_4/CO_2)$ should be less than −4.0 in high-temperature, mantle-derived gas[23], and if CH_4 were present then $\delta^{13}C$ CH_4 would be above −25. Errors for Yellowstone gases are smaller than the symbol size. Data are from ref. 10.

Table 1 | CO_2, He and radiogenic He flux for Yellowstone and select thermal areas

Region	Q_{CO_2} (t d^{-1})	Q_{He} (mol yr^{-1})	R/R_a	$q^C_{He}(1)$ (mol yr^{-1})	$q^C_{He}(2)$ (mol yr^{-1})	WC(1)	WC(2)
Heart Lake Basin	6	6.91×10^3	3.0	5.97×10^3	5.03×10^3	0.35	0.29
Brimstone Basin	277	2.50×10^5	3.0	2.16×10^5	1.82×10^5	12.6	10.6
Hot Spring Basin	410	2.39×10^5	5.5	1.79×10^5	1.19×10^5	10.5	7.0
Roaring Mountain	100	8.49×10^4	1.9	7.77×10^4	7.05×10^4	4.5	4.1
Entire park (min)	20,000	3.95×10^6	6.9	2.71×10^6	1.47×10^6	159	86.2
Entire park (max)	80,000	4.92×10^7	6.9	3.38×10^7	1.83×10^7	1,970	1,070
Entire park (best)	45,000	1.49×10^7	6.9	1.02×10^7	5.55×10^6	597	324

$q^C_{He}(1)$ is the crustal He flux assuming a mantle R/R_a of 22. $q^C_{He}(2)$ is the crustal He flux assuming a mantle R/R_a of 11. WC(1) and WC(2) are the two corresponding calculated crustal ^{4}He fluxes relative to the ^{4}He annually produced by a 43-km-thick crustal section (9,000 km^2). See the Supplementary Information for the source data for Table 1.

eastern edge, older rocks almost certainly make up the mid to lower crust that hosts the intrusive complex feeding the surface volcanism[24]. Yellowstone sits astride the Wyoming craton near the northeast–southwest-trending boundary between two of its subdivisions, the Montana sedimentary province and the Beartooth–Bighorn magmatic zone[25]. The former has rocks as old as 3.5 Gyr with detrital zircons as old as 4.0 Gyr. Early Yellowstone rhyolites inherited a strong isotope signature (in Sr, Nd and Pb) from this old crust, implying that the rhyolites were formed through hybridization of fractionates of mantle-derived basalts with melts of crustal Archaean rock[26]. More recent volcanism, particularly within the caldera, indicates less input from old crust, because this fertile source wanes in abundance and new magma is generated by recycling of hydrothermally altered volcanic rocks and fractionation of the existing magma reservoir[24,26,27]. Only outside the caldera margins does a strong Archaean signature remain[26]. Similarly, it is likely that early magmatism caused greater amounts of crustal degassing within the caldera, and that the present measured flux of crustal ^{4}He is somewhat reduced relative to the initial stages of volcanism.

Numerous studies indicate that this part of the Wyoming craton was stabilized by 2.8 Gyr ago and since then has undergone relatively little tectonism until the present day. Potassium–argon ages of ~2.3 Gyr from cratonic basement rocks imply that they stayed at a temperature of less than ~350 °C for more than 2 Gyr, and that development of a stable tectosphere and deep mantle root shielded the area from subsequent regional orogens[25,28]. Consequently, there has been relatively little potential to mobilize accumulated He.

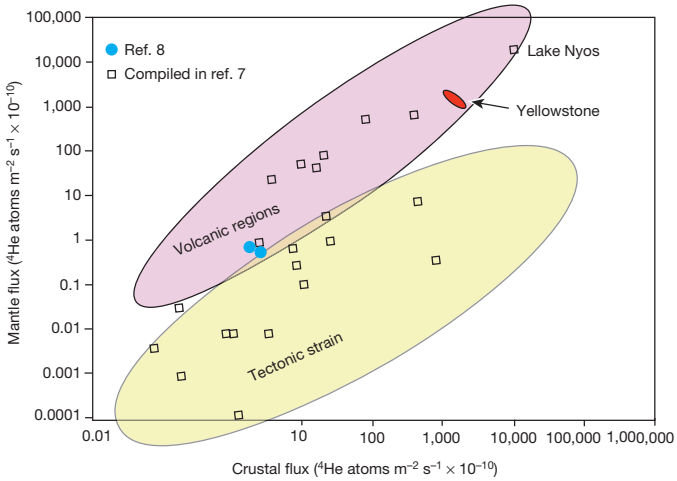


Figure 3 | Crustal versus mantle ^{4}He flux based on ref. 7. Data for the original diagram were modelled assuming that $R \approx 0.01R_a$ in the crust and $R = 8.6R_a$ in the mantle. Volcanic regions were found to release significantly more mantle He, but also frequently more crustal He, than non-volcanic areas. The Yellowstone data lie at the high-flux end of the volcanic field for both mantle and crustal He. Values for the Yellowstone field cover the region between the two mantle endmembers used in Table 1. Data from ref. 8 represent an early study using similar methods, but in a younger, less (volcanically) active crustal region. Lake Nyos is a small crater lake in Cameroon with abundant CO_2 discharge.

The elements U and Th, which dominate the production of radiogenic He in the crust[19], are located primarily in the mineral zircon and other accessory minerals such as monazite, titanite and apatite that are common in igneous and metamorphic rocks. Above a mineral's theoretical blocking temperature, an element of interest will diffuse out of the mineral and into the surrounding rocks. Helium should diffuse out of zircon and titanite at temperatures exceeding ~200 °C (ref. 29), as would occur at depths >10 km in all but the lowest of geothermal gradients. However, without a path for advective fluid flow in the host rocks, He cannot escape and will remain trapped in the very low-permeability rocks that reside in undisturbed (that is, tectonically unmodified) mid-to-lower crust[30]. Without significant fluid advection or large He concentration gradients, the rate of He diffusion is simply too slow to allow for He migration on crustal timescales[19].

Examples of long-term He residence in the crust are reported increasingly often. Recently, Archaean waters were identified in fractures in a mine from the Precambrian Shield of Ontario[31]. A residence time of 1.1 ± 0.6 Gyr was calculated for He in this fluid. In many rocks, fluid inclusions are ideal receptacles for He (ref. 32), which partitions strongly into any gas bubble relative to the host mineral: fluid inclusions in the Archaean Witwatersrand Basin, in South Africa, retained their initial trapped Ne for much more than 2 Gyr (ref. 33). And over the past few decades, workers have become increasingly convinced that both mantle and radiogenic He can accumulate over geological time within crustal aquifers and gas reservoirs[3,4,6,7,19].

Never before, though, has non-steady-state emission of radiogenic gas been so clearly illustrated. Over the past 2 Myr, the Yellowstone hotspot has penetrated one of the oldest cratons on the planet, resulting in tectonic strain, fracturing, melting and infiltration by circulating metamorphic and hydrothermal fluids. About 50% of our samples have $CO_2/^3$He ratios between 2×10^9 and 4×10^9, implying that a consistent mantle signature for CO_2 is present throughout the region. The mantle-derived gas flux scavenges any crustal ^{4}He made increasingly available through the tectonic and metamorphic processes noted above. A recent compilation of values for crustal He flux noted increased crustal and mantle He flux in volcanic regions[7], relative to non-volcanic terrains, but the data set was largely limited to crater lakes a few square kilometres in size and for time periods of years to decades. Our Yellowstone results fall on the upper part of this trend and greatly expand the spatial and temporal scales (Fig. 3).

Noble-gas data are frequently interpreted by assuming closed systems, where circulating fluids interact solely with rocks and sediments of local aquifers. More rapid addition of crustal ^{4}He implies that mean groundwater transit times can be orders of magnitude shorter than those calculated on the basis of closed-system assumptions. Our work demonstrates that open-system behaviour can dominate noble-gas budgets in regions of active magmatism and tectonics. Yet it also demonstrates an example of closed-system behaviour within a stable Archaean craton, where He initially accumulated for billions of years before incursion of the Yellowstone hotspot.

METHODS SUMMARY

Gas samples were collected from fumaroles and hot springs in evacuated, NaOH-bearing glass bottles according to standard methods[10]. Diffuse CO_2 flux was estimated by the accumulation chamber technique[11–13,16]. Gas chemistry was determined with a

combination of gas chromatography, ion chromatograph and manometry[10] in the geothermal laboratory in Menlo Park, California. Noble-gas isotopic ratios, including those for He, were measured in Denver, Colorado, at the USGS Noble Gas Laboratory. A detailed description of the methods and further information about the assumptions used in our calculations are provided in Methods.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 August; accepted 11 December 2013.

1. Porcelli, D., Ballentine, C. J. & Wieler, R. *Noble Gases in Geochemistry and Cosmochemistry* (Rev. Mineral. Geochem. 47, Mineralogical Society of America and the Geochemical Society, 2002).
2. Craig, H., Lupton, J. E., Welhan, J. A. & Poreda, R. Helium isotope ratios in Yellowstone and Lassen Park volcanic gases. *Geophys. Res. Lett.* **5,** 897–900 (1978).
3. Ballentine, C. J., Burgess, R. & Marty, B. Tracing fluid origin, transport and interaction in the crust. *Rev. Mineral. Geochem.* **47,** 539–614 (2002).
4. Ballentine, C. J. & Sherwood Lollar, B. Regional groundwater focusing of nitrogen and noble gases in the Hugoton-Panhandle giant gas field, USA. *Geochim. Cosmochim. Acta* **66,** 2483–2497 (2002).
5. Kennedy, B. M. *et al.* Mantle fluids in the San Andreas fault system, California. *Science* **278,** 1278–1281 (1997).
6. Torgersen, T. & Clarke, W. B. Helium accumulation in groundwater, I: an evaluation of sources and the continental flux of crustal ^{4}He in the Great Artesian Basin, Australia. *Geochim. Cosmochim. Acta* **49,** 1211–1218 (1985).
7. Torgersen, T. Continental degassing flux of ^{4}He and its variability. *Geochem. Geophys. Geosyst.* **11,** Q06002 (2010).
8. Sano, Y., Wakita, H. & Huang, C.-W. Helium flux in a continental land area estimated from ^{3}He/^{4}He ratio in northern Taiwan. *Nature* **323,** 55–57 (1986).
9. Kennedy, B. M., Lynch, M. A., Reynolds, J. H. & Smith, S. P. Intensive sampling of noble gases in fluids at Yellowstone, I. Early overview of the data; regional patterns. *Geochim. Cosmochim. Acta* **49,** 1251–1261 (1985).
10. Bergfeld, D. B. *et al. Gas and Isotope Chemistry of Thermal Features in Yellowstone National Park, Wyoming*. Scientific Investigations Report 2011–5012 (US Geological Survey, 2011).
11. Werner, C. *et al.* Volatile emissions and gas geochemistry of Hot Spring Basin, Yellowstone National Park, USA. *J. Volcanol. Geotherm. Res.* **178,** 751–762 (2008).
12. Lowenstern, J. B., Bergfeld, D., Evans, W. C. & Hurwitz, S. Generation and evolution of hydrothermal fluids at Yellowstone: insights from the Heart Lake Geyser Basin. *Geochem. Geophys. Geosyst.* **13,** Q01017 (2012).
13. Bergfeld, D. B., Evans, W. C., Lowenstern, J. B. & Hurwitz, S. Carbon dioxide and hydrogen sulfide degassing and cryptic thermal input to Brimstone Basin, Yellowstone National park, Wyoming. *Chem. Geol.* **330–331,** 233–243 (2012).
14. Smith, R. B. *et al.* Geodynamics of the Yellowstone hotspot and mantle plume: Seismic and GPS imaging, kinematics, and mantle flow. *J. Volcanol. Geotherm. Res.* **188,** 26–56 (2009).
15. Lowenstern, J. B. & Hurwitz, S. Monitoring a supervolcano in repose: heat and volatile flux at the Yellowstone Caldera. *Elements* **4,** 35–40 (2008).
16. Werner, C. & Brantley, S. CO_2 emissions from the Yellowstone volcanic system. *Geochem. Geophys. Geosyst.* **4,** 1061 (2003).
17. Lorenson, T. D. & Kvenvolden, K. A. in *The Future of Energy Gases* (eds Howell D. G. *et al.*) 453–470 (Profession Paper 1570, US Geological Survey, 1993).
18. Chiodini, G. *et al.* Insights from fumarole gas geochemistry on the origin of hydrothermal fluids on the Yellowstone Plateau. *Geochim. Cosmochim. Acta* **89,** 265–278 (2012).
19. Ballentine, C. J. & Burnard, P. G. Production, release and transport of noble gases in the continental crust. *Rev. Mineral. Geochem.* **47,** 481–538 (2002).
20. Rudnick, R. & Fountain, D. M. Nature and composition of the continental crust: a lower crustal perspective. *Rev. Geophys.* **33,** 267–309 (1995).
21. Vaughan, R. G., Keszthelyi, L. P., Lowenstern, J. B., Jaworowski, C. & Heasler, H. Use of ASTER and MODIS thermal infrared data to quantify heat flow and hydrothermal change at Yellowstone National Park. *J. Volcanol. Geotherm. Res.* **233–234,** 72–89 (2012).
22. Graham, D. W. *et al.* Mantle source provinces beneath the Northwestern USA delimited by helium isotopes in young basalts. *J. Volcanol. Geotherm. Res.* **188,** 128–140 (2009).
23. Chiodini, G. CO_2/CH_4 ratio in fumaroles a powerful tool to detect magma degassing at quiescent volcanoes. *Geophys. Res. Lett.* **36,** L02302 (2009).
24. Christiansen, R. L. *The Quaternary and Pliocene Yellowstone Plateau Volcanic Field of Wyoming, Idaho, and Montana.* Professional Paper 729-G (US Geological Survey, 2001).
25. Mueller, P. & Frost, C. The Wyoming Province: a distinctive Archean craton in Laurentian North America. *Can. J. Earth Sci.* **43,** 1391–1397 (2006).
26. Hildreth, W., Halliday, A. N. & Christiansen, R. L. Isotopic and chemical evidence concerning the genesis and contamination of basaltic and rhyolitic magma beneath the Yellowstone Plateau Volcanic Field. *J. Petrol.* **32,** 63–138 (1991).
27. Bindeman, I. N., Fu, B., Kita, N. T. & Valley, J. W. Origin and evolution of silicic magmatism at Yellowstone based on ion microprobe analysis of isotopically zoned zircons. *J. Petrol.* **49,** 163–193 (2008).
28. Mueller, P. A., Mogk, D. W., Henry, D. J., Wooden, J. L. & Foster, D. A. Geologic evolution of the Beartooth Mountains: insights from petrology and geochemistry. *Northwest Geol.* **37,** 5–20 (2008).
29. Farley, K. A. (U-Th)/He dating: techniques, calibrations and applications. *Rev. Mineral. Geochem.* **47,** 819–844 (2002).
30. Ingebritsen, S. E. & Manning, C. E. Permeability of the continental crust: dynamic variations from seismicity and metamorphism. *Geofluids* **10,** 193–205 (2010).
31. Holland, G. *et al.* Deep fracture fluids isolated in the crust since the Precambrian era. *Nature* **497,** 357–360 (2013).
32. Sapienza, G., Hilton, D. R. & Scribano, V. Helium isotopes in peridotite mineral phases from Hyblean Plateau xenoliths (south-eastern Sicily, Italy). *Chem. Geol.* **219,** 115–129 (2005).
33. Lippmann-Pipke, J. *et al.* Neon identifies two billion year old fluid component in Kappvaal Craton. *Chem. Geol.* **283,** 287–296 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank S. Ingebritsen for a review. C. Hendrix, S. Gunther, H. Heasler and D. Mahony assisted with field planning and logistics.

Author Contributions J.B.L. and D.B. together led the sampling program at Yellowstone. W.C.E. participated in the fieldwork. D.B. did all laboratory analyses except the noble-gas analyses. D.B. also led the diffuse degassing fieldwork at Brimstone Basin and Heart Lake. A.G.H. performed the noble-gas analyses. J.B.L. first recognized the significance of the He emission rates for crustal degassing and wrote the first draft of the manuscript. W.C.E. provided considerable input on gas geochemistry and noble-gas systematics, and assisted with subsequent drafts. All authors edited later versions of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.B.L. (jlwnstrn@usgs.gov).

LETTER

doi:10.1038/nature12874

Species coexistence and the dynamics of phenotypic evolution in adaptive radiation

Joseph A. Tobias[1]*, Charlie K. Cornwallis[1,2]*, Elizabeth P. Derryberry[3,4], Santiago Claramunt[3,5,6], Robb T. Brumfield[3,5] & Nathalie Seddon[1]

Interactions between species can promote evolutionary divergence of ecological traits and social signals[1,2], a process widely assumed to generate species differences in adaptive radiation[3–5]. However, an alternative view is that lineages typically interact when relatively old[6], by which time selection for divergence is weak[7,8] and potentially exceeded by convergent selection acting on traits mediating interspecific competition[9]. Few studies have tested these contrasting predictions across large radiations, or by controlling for evolutionary time. Thus the role of species interactions in driving broad-scale patterns of trait divergence is unclear[10]. Here we use phylogenetic estimates of divergence times to show that increased trait differences among coexisting lineages of ovenbirds (Furnariidae) are explained by their greater evolutionary age in relation to non-interacting lineages, and that—when these temporal biases are accounted for—the only significant effect of coexistence is convergence in a social signal (song). Our results conflict with the conventional view that coexistence promotes trait divergence among co-occurring organisms at macroevolutionary scales, and instead provide evidence that species interactions can drive phenotypic convergence across entire radiations, a pattern generally concealed by biases in age.

Phenotypic divergence through species interaction is one of the oldest concepts in evolutionary biology. The underlying mechanism—first called 'divergence of character'[3], now generally known as character displacement[11]—is compellingly simple: divergent phenotypes should be favoured when closely related species interact to minimize the costs of ecological competition, misdirected aggression or hybridization[12]. This deterministic mechanism of selection, acting on the ecological and social traits of individuals and operating simultaneously across species, is believed to contribute to pervasive macroevolutionary patterns, including the tree-like structure of trait divergence over entire radiations[10,13,14] and the non-random morphological differences found almost universally among co-occurring lineages[14,15]. However, although the importance of character displacement as a microevolutionary process is well documented in pairs or small numbers of species[1,2,16], the broader implications for macroevolution are uncertain.

The main problem is that different processes can result in similar broad-scale patterns of phenotypic variation. When viewed across species, the key prediction of character displacement is that trait differences will be greater in coexisting (that is, sympatric) than non-interacting (allopatric) lineages[14]. However, because allopatric speciation is the norm, lineages may already be ancient by the time they interact in sympatry, particularly if range overlap is constrained by competitive exclusion[17]. Thus, greater phenotypic divergence among sympatric species may simply reflect trait differences acquired in allopatry, and accentuated by the 'ecological sorting' of pre-existing phenotypes[18,19]. Moreover, this biogeographical pattern could also obscure the adaptive convergence of interacting competitors[9].

These opposing hypotheses can only be tested by assessing phenotypic differences among related lineages in the context both of geographical space and of evolutionary time[12,17]. Such tests are required across broad samples of species to ensure that results do not reflect chance events or special cases. This is particularly critical given that most studies demonstrating character displacement have focused on young or species-poor systems, such as lakes and archipelagos, characterized by early sympatry[12]. The most immediate priority is to quantify phenotypic divergence in a spatial and temporal framework, and across numerous lineages with comparable functional traits, including both resource-exploiting traits and social signals. This has not been achieved so far because of the challenges posed by sampling across extensive radiations.

To address this issue, we examined trait divergence in relation to interactions among 350 lineages of ovenbirds, a radiation of tracheophone suboscine birds that has evolved into a remarkable variety of phenotypes over the past 35 Myr (Extended Data Fig. 1). We estimated divergence in three key functional traits associated with competition (beaks), locomotion (tarsi) and social interaction (songs) (Fig. 1). The beak is tightly linked to resource acquisition[15]; tarsus length provides an independent index of foraging niche and body size[20]; and songs provide insight into both reproductive and agonistic interactions because they function in mate attraction and territoriality[2,9]. Importantly, all lineages of ovenbirds share the same basic ecological niche (insectivory), and their songs, unlike those of most passerine birds, are structurally simple and apparently innate (Extended Data Figs 2 and 3 and Supplementary Information).

Comparing divergence between each ovenbird lineage and its closest relative in sympatry ($n = 270$) and allopatry ($n = 249$) showed that sympatric lineages have undergone significantly greater divergence than allopatric lineages in beaks (linear mixed model (LMM): $F_{1,305} = 49.65$; $P < 0.001$), tarsi ($F_{1,302} = 30.95$; $P < 0.001$) and songs ($F_{1,285} = 6.90$; $P = 0.009$) (Fig. 1c–e and Supplementary Table 3). Similar findings are widespread and often interpreted as evidence for interspecific competition or character displacement (see, for example, refs 14, 21). However, it is possible that these differences between sympatry and allopatry are simply caused by disparities in evolutionary age, as phylogenetic data (Fig. 2) indicate that closest relatives in sympatry are on average 2.3 times older than those in allopatry ($F_{1,290} = 286.1$; $P < 0.0001$) (Supplementary Table 3, Fig. 1f and Extended Data Fig. 4). As is typical for vertebrates, ovenbird lineages therefore undergo an extended allopatric phase preceding secondary contact, a temporal pattern often proposed to reflect competitive exclusion among lineages with conserved ecological niches[6,7,19].

When we controlled for this difference in age between sympatric and allopatric lineages, the association between coexistence and divergence was removed (Fig. 3). There was no significant effect of sympatry on divergence in tarsi (phylogenetic linear mixed model (PLMM): $F_{1,14} = 0.01$; $P = 0.93$; Supplementary Table 4) or beaks ($F_{1,39} = 0.33$; $P = 0.57$; Supplementary Table 5), and the only pattern consistently detected was an increased similarity of songs in sympatry ($F_{1,33} = 5.95$; $P = 0.02$; Supplementary Table 6). These results were robust to best-fit models of

[1]Edward Grey Institute, Department of Zoology, University of Oxford, Oxford OX1 3PS, UK. [2]Department of Biology, Lund University, Lund, SE-223 62, Sweden. [3]Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803, USA. [4]Department of Ecology and Evolutionary Biology, Tulane University, New Orleans, Louisiana 70118, USA. [5]Museum of Natural Science, Louisiana State University, Baton Rouge, Louisiana 70803, USA. [6]Department of Ornithology, American Museum of Natural History, New York, New York 10024, USA.
*These authors contributed equally to this study.

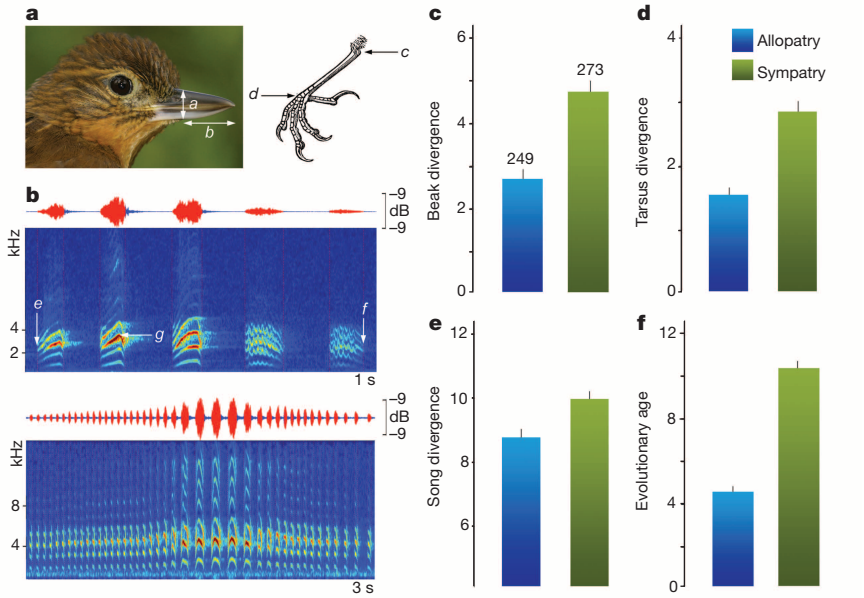


Figure 1 | Phenotypic divergence and evolutionary age in co-occurring lineages. **a**, Ecological traits of ovenbirds (beak depth (*a*) and length (*b*); tarsus length (*c* to *d*)). **b**, Social traits (spectrograms (frequency/time) and waveforms (amplitude/time) of typical songs show duration (*e* to *f*), pace (*e* to *f*/number of notes) and peak frequency (*g*)). Closest relatives were more divergent in sympatry than allopatry in (**c**) beak (LMM: $P < 0.001$), (**d**) tarsus ($P < 0.001$), (**e**) song ($P = 0.009$) and (**f**) evolutionary age (length of time elapsed since sharing a common ancestor ($P < 0.001$)). Units of measurement: **c**, PC1 scores derived from a phylogenetic principal components analysis on three beak variables; **d**, mm; **e**, distance between centroids derived from a phylogenetic principal components analysis of all acoustic traits for song; **f**, Myr, calculated from mitochondrial DNA sequence divergence. Bars, mean level of divergence ± s.d.; sample sizes are given in **c**. (Photograph (*Syndactyla striata*) and spectrograms (upper, *Synallaxis erythrothorax*; lower, *Cinclodes aricomae*) by J.A.T.).

trait evolution and divergence, including accounting for different models of trait evolution in sympatry versus allopatry, as well as the bounded evolution of song (Supplementary Tables 7–10). The same patterns also held when we focused on species-level taxa by removing intraspecific lineages (Supplementary Tables 11–13), and when we restricted data sets exclusively to pairs of sister species ($n = 111$) (Supplementary Table 14).

One possibility is that the classic pattern of character displacement is not detected by these analyses because it is confined to the youngest lineages, as hinted by apparent greater divergence in sympatric lineages during the initial 6 Myr after speciation (Fig. 3a, b). To test this, we re-ran our models excluding all comparisons between lineages >6 Myr old. This analysis showed no divergence in sympatry for any trait, and removed significant convergence in songs (Supplementary Tables 16–18),

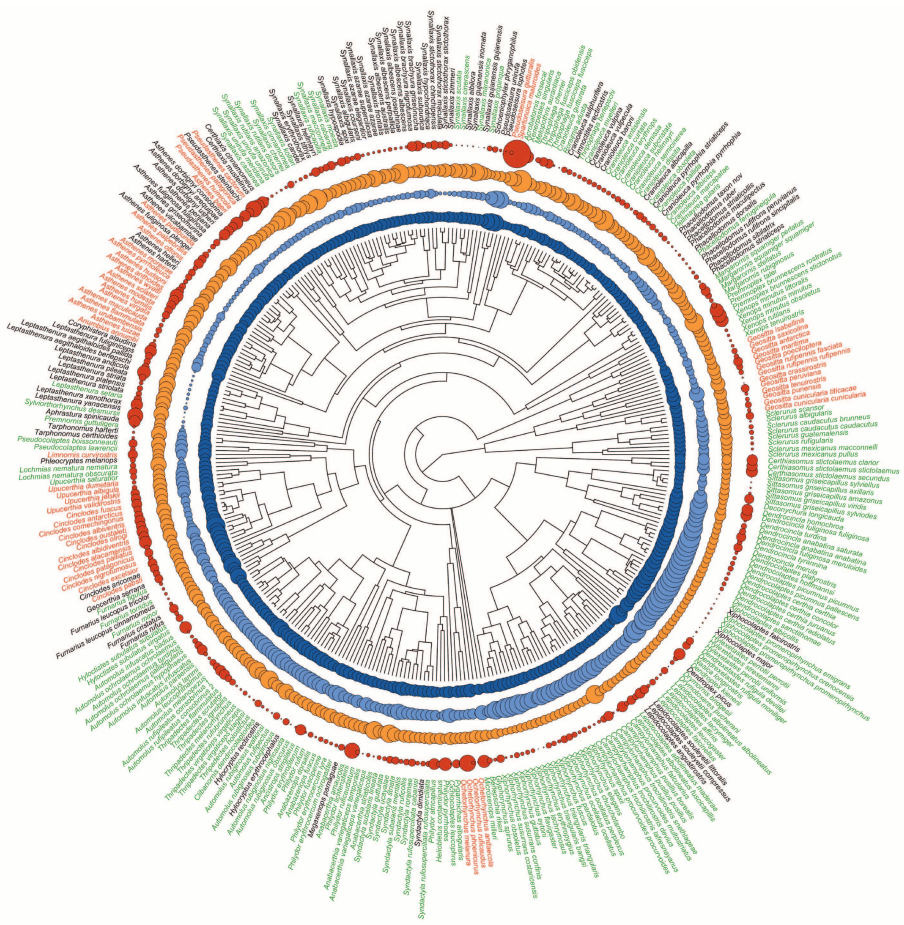


Figure 2 | Phenotype in relation to habitat and evolutionary history. The phylogram shows associations between habitat and phenotypic traits across 350 ovenbird lineages. Names are colour-coded by primary habitat structure: green, dense; black, semi-open; red, open. Dots represent variation in ecological traits (dark blue, tarsus length; light blue, beak morphology) and social traits (orange, song peak frequency; red, song pace). Dot size represents phenotypic variation (large dots, longer tarsi, larger beaks, higher peak and faster pace).

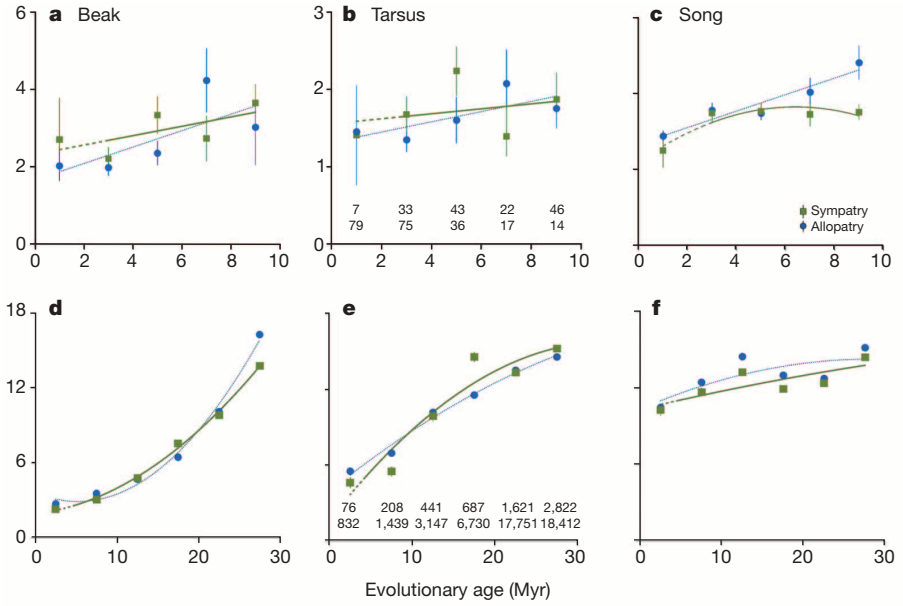


Figure 3 | Comparing divergence in sympatry versus allopatry. Divergence in ovenbird beaks (**a**, **d**), tarsi (**b**, **e**) and songs (**c**, **f**) over time, calculated as Euclidean distances (units in Fig. 2) between multiple pairs of lineages. Upper panels, mean trait differences for closest relatives (sample sizes in **b**); lower panels, data for all unique lineage combinations (sample sizes in **e**). Curves are fitted for significant quadratic relationships (PLMM: $P < 0.05$). Dashed lines, small samples. Error bars, s.d. (very small in **d**–**f**).

suggesting that (1) we had not overlooked ecological or reproductive character displacement simply because they only occur in the early stages of divergence, and (2) the similarity of songs in sympatry is reduced in the youngest lineages, perhaps because the increased risk of hybridization impedes signal convergence in recently diverged species[8].

To evaluate the effect of coexistence over longer timeframes, we modelled all pairwise comparisons ($n = 34{,}588$ pairs of lineages), again showing no significant signature of ecological character displacement when controlling for evolutionary age (Fig. 3d, e and Supplementary Tables 19 and 20). Indeed, beaks were significantly more similar in sympatry than allopatry, although this effect was weak (PLMM: $F_{1,33638} = 6.04$; $P = 0.01$) and largely driven by numerous ancient sympatric species (Extended Data Fig. 5) in which beaks may be shaped by convergent adaptation to shared ecological niches[20] (for example, terrestrial lineages in clades 8 (Furnariini) and 12 (Sclerurinae); see Extended Data Fig. 1). Similar environmental factors may contribute to song convergence at this broad taxonomic scale as we found positive relationships between song and habitat divergence, in line with the acoustic adaptation of songs to habitats with different transmission properties[22], and between song and beak morphology, consistent with correlated evolution between ecological and social traits[23] (Supplementary Table 21). Nonetheless, even after controlling for these interactions, the songs of sympatric lineages were more similar than those of allopatric lineages, an effect that was both relatively strong ($F_{1,33994} = 27.13$; $P < 0.0001$; see Extended Data Fig. 4) and apparently consistent regardless of the evolutionary age of interacting lineages (Fig. 3f). To verify these findings, we conducted a series of simulation tests that confirmed our results were not explained by the structure of our data or the distribution of trait divergences (Extended Data Fig. 6), or by variation in the mode of trait evolution (Extended Data Figs 7 and 8).

The preceding analyses treat sympatry and allopatry as binary variables to facilitate interpretation, in line with most previous work on character displacement. However, as the cut-off between these geographical states is somewhat arbitrary, we re-ran models including proportional range overlap as a covariate (Supplementary Tables 22–27 and Fig. 4). Focusing on closest relatives, we found no effect of range overlap on divergence in beaks (Fig. 4d) or tarsi (Fig. 4e), again rejecting the central prediction of character displacement theory. In contrast, there was a strong positive relationship between song similarity and range overlap (Fig. 4f), both when considering interactions among closest relatives ($F_{1,304} = 10.04$; $P = 0.002$; Supplementary Table 24), or all lineages ($F_{1,33962} = 23.13$; $P < 0.0001$; Fig. 4l and Supplementary Table 27).

Regardless of how sympatry was measured, we found that phenotypic divergence is best predicted by evolutionary age, suggesting that most trait differences among lineages simply accumulate over evolutionary time as a result of processes such as genetic drift and ecological adaptation. This may explain why previous studies have reported the signature of ecological and reproductive character displacement to be absent[24,25] or equivocal[17] at the scale of clades and communities. However, it is striking that we find no evidence of divergent character displacement in bird beaks and songs, two traits that have played a central role in the development of character displacement theory[1,2,15]. Our phylogenetic comparative approach may have obscured individual cases of character displacement, yet our results indicate that this process is subtle or infrequent in ovenbirds, and fails to predict broad-scale patterns. We can rule out the possibility that our conclusions are specific to clades with weak species interactions as ovenbirds are generally territorial and predisposed to high levels of interspecific competition[26]. It is also unlikely that we have overlooked displacement because of statistical issues as our models have sufficient power to deal with noisy data, and even detect a signature of increased song similarity in sympatry.

This apparent convergence in songs is counter-intuitive from the perspective of classic character displacement theory, and opposite to the patterns generally assumed to be pervasive in birds[2,27]. Yet, it is consistent with the view that agonistic character displacement can drive adaptive convergence of signals mediating competitor recognition in multispecies systems[9]. This hypothesis is based on the idea that individuals with convergent agonistic signals have higher fitness because they are better at defending resources against both conspecific and heterospecific competitors. In birds, competition for food or territories can take place among relatively old sympatric lineages with partly overlapping ecological niches[27]. Although agonistic character displacement has rarely been demonstrated, our findings align with previous studies suggesting that social signals may converge owing to competition[9,27,28], and that hybridization is then averted because receivers adapt to differentiate similar signals[8].

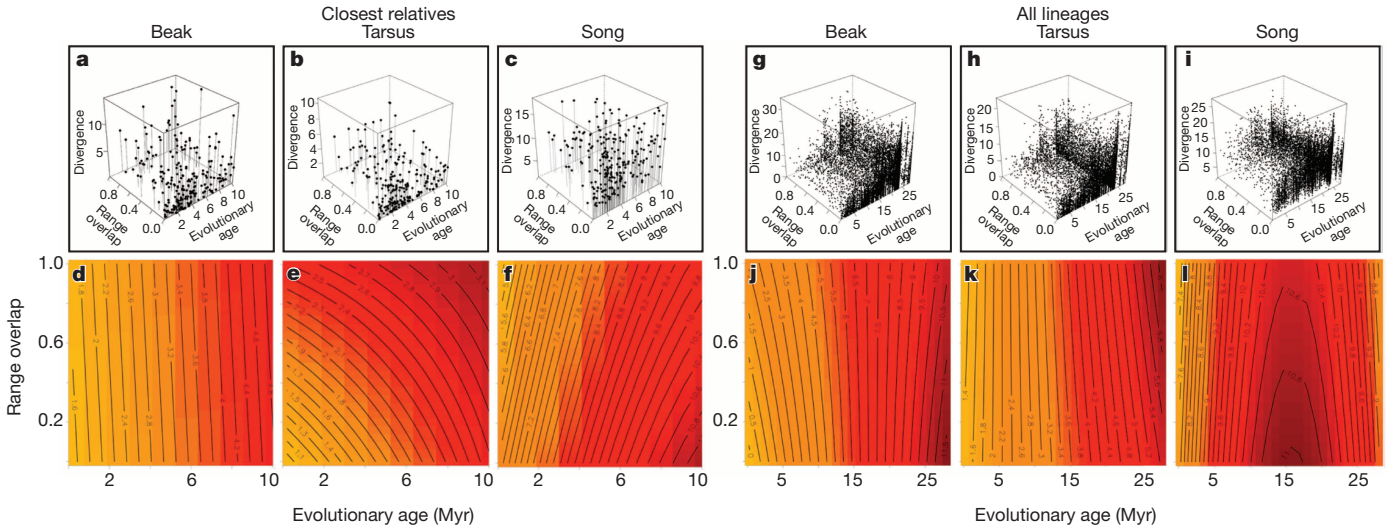


Figure 4 | The dynamics of phenotypic divergence across space and time. Divergence in ovenbird beaks, tarsi and songs between closest relatives (**a**–**f**) and all lineages (**g**–**l**) plotted as a function of proportional geographical range overlap and evolutionary age (Myr). Three-dimensional scatter plots of trait divergence (**a**–**c**; **g**–**i**) show that data are noisy, as is expected with phenotypic trait evolution across large radiations. Because speciation tends to occur in allopatry, sampling of young, broadly sympatric lineages is relatively sparse (hence our preference for treating sympatry and allopatry as binary variables in models). Note that **g**–**i** contain pairwise comparisons among all lineages and thus species contribute information to multiple data points. Heatmaps (**d**–**f**; **j**–**l**) are shown to clarify the effect of proportional range overlap (*y* axis) and evolutionary time (*x* axis) on trait divergence (darker colours indicate greater divergence). For trait units, see Fig. 1c–e; for sample sizes, see Fig. 3.

An alternative explanation is that widespread hybridization occurs in ovenbirds, causing the introgression of genes coding for song structure. This seems unlikely given that hybridization should reduce molecular sequence divergence in sympatry, lowering the estimated age of interacting lineages (cf. Extended Data Fig. 4). Another possibility is that such lineages occur in more similar environments, potentially driving convergence through ecological adaptation. However, using coarse habitat categories, we found no significant difference in habitat divergence between sympatric and allopatric pairs of lineages (PLMM: $F_{5,258} = 1.59$, $P = 0.16$), largely because habitat niche divergence only occurs after an extended time-lag, regardless of geographical relationships (Extended Data Fig. 5). Indeed, previous studies have shown that even narrowly defined microhabitat niches are conserved over millions of years in ovenbirds, during which time sympatry is associated with niche divergence[26]. Given that a pattern of greater song similarity in sympatry is detected throughout this initial period of divergence (Fig. 4f), our results cannot simply be explained by acoustic adaptation to shared habitat. A related concern is that song divergence may be more bounded in sympatry than allopatry, generating spurious evidence of convergence. This may occur, for example, if sympatric ovenbirds tend to coexist in dense habitat where song divergence is potentially limited by constraints on signal transmission[29]. However, we found that song divergence was bounded regardless of geographical relationship, and although the constraint parameter (α) differed marginally in sympatry and allopatry, a strong signature of convergence was retained even after accounting for this discrepancy (Supplementary Table 10).

This signature may be produced by classic convergence (that is, decreasing trait differences over time), or else simply a failure to diverge. Although it is difficult to discriminate between these outcomes, both can be viewed as forms of convergence (Supplementary Information), and it seems likely that classic convergence contributes to the pattern. The hump-shaped trajectory of song divergence among closest relatives in sympatry (Fig. 3c), as indicated by a significant negative quadratic term in our models (Supplementary Tables 6 and 9), is not predicted by constraints on divergence. It is more consistent with accentuated convergence among older lineages owing to the accumulation of sympatry over time—that is, the average duration of coexistence increases gradually with age[26]. Furthermore, we assume that normal levels of divergence occur during the allopatric phase after speciation in ovenbirds, and thus patterns of reduced divergence in sympatry are most probably produced by classic convergence after secondary contact. This possibility requires further testing with detailed field studies.

Although we do not measure species interactions directly, they offer the most likely explanation for song convergence in ovenbirds for three reasons. First, interactions among sympatric lineages are widespread, and often cause shifts in trait evolution[5,12]. Second, the contrast between non-converging beaks and tarsi (Fig. 4d, e) and converging songs (Fig. 4f) suggests that opposing mechanisms regulate divergence in ecological versus social traits. Given that ecological trait divergence in ovenbirds is explained by adaptation to microhabitat or foraging substrate[20], the implication is that song convergence may be driven by a separate, socially mediated mechanism. Third, the existence of remarkably non-divergent songs among interspecifically territorial ovenbird lineages several million years after speciation (Extended Data Fig. 9) is unlikely to arise through learning, hybridization, acoustic adaptation or bounded evolution, and yet is consistent with agonistic interactions[9,27,28].

We have shown that uncorrected patterns of phenotypic variation are confounded by biases in the time available for trait differences to evolve, and exaggerate the role of character displacement at macroevolutionary scales. Once these biases are taken into account, we find no evidence that trait differences in coexisting ovenbirds are explained by species interactions, even among the 'most closely-allied forms' that Darwin[3] predicted would diverge as a result of 'the severest competition'. We propose that ecological and reproductive character displacement is restricted to cases of early secondary contact, particularly involving incipient species, and that this phenomenon is comparatively rare in the ancient, continental radiations that make up much of biodiversity.

Overall, our analyses support the hypothesis that lineages build up enough ecological and reproductive isolation during allopatry to bypass character displacement after secondary contact[7,19,26]. These results therefore challenge the assumption that ecological and reproductive character displacement are key microevolutionary mechanisms contributing to macroevolutionary patterns, including trait differences in communities and across adaptive radiations[4,10,14]. Moreover, we have shown that the same biases in evolutionary age that explain character differences in sympatric lineages potentially mask the role of convergent evolution.

A pervasive pattern of phenotypic convergence is not new, even in radiations[30], but our findings provide the first evidence that such outcomes can be driven by species interactions. We conclude that macroevolutionary patterns previously interpreted as character displacement should be re-analysed in an explicitly temporal framework, as doing so may show that mechanisms of character convergence are widespread.

METHODS SUMMARY

Focusing on 350 ovenbird lineages, we measured ecological traits (beak shape, tarsus length) from museum specimens, and a social trait (song structure) from digital recordings (see Fig. 1). Habitat preferences were scored from the literature. We used digital range polygons to quantify geographical range overlap as a continuous variable, and used a 20% range overlap threshold to convert this into a binary variable (sympatry/allopatry). To estimate the evolutionary time since divergence for all pairs of lineages, we produced a phylogenetic tree (Fig. 2) using standard techniques[6,20]. We then used PLMMs to compare phenotypic divergence in sympatry versus allopatry, controlling for shared ancestry, evolutionary age and habitat differences. We focused at two taxonomic levels: closest relatives and the entire radiation. To assess whether our results were robust to different models of trait evolution, we also used phylogenetic generalized least squares (PGLS), and verified a new approach for accounting for best-fit models in both sympatry and allopatry using PLMMs. We examined the sensitivity of the results to several refinements, including restricting analyses to young lineages (<6 Myr) or sister species, and treating proportional range overlap as a continuous variable. Data manipulation was conducted in *R* version 3.0.1, PLMMs in ASReml-R, and PGLS analyses used R libraries 'ape' and 'nlme' (see Methods for all software references). For discussion and definition of the terms 'adaptive radiation', 'character displacement' and 'convergence' see Supplementary Information.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 March; accepted 8 November 2013.
Published online 22 December 2013.

1. Grant, P. R. & Grant, B. R. Evolution of character displacement in Darwin's finches. *Science* **313,** 224–226 (2006).
2. Grant, B. R. & Grant, P. R. Songs of Darwin's finches diverge when a new species enters the community. *Proc. Natl Acad. Sci. USA* **107,** 20156–20163 (2010).
3. Darwin, C. R. *On the Origin of Species* (John Murray, 1859).
4. Simpson, G. G. *Tempo and Mode in Evolution* (Columbia Univ. Press, 1944).
5. Pfennig, D. W. & Pfennig, K. S. Character displacement and the origins of diversity. *Am. Nat.* **176,** S26–S44 (2010).
6. Weir, J. T. & Price, T. D. Limits to speciation inferred from times to secondary sympatry and ages of hybridizing species along a latitudinal gradient. *Am. Nat.* **177,** 462–469 (2011).
7. Price, T. D. The roles of time and ecology in the continental radiation of the Old World leaf warblers (*Phylloscopus* and *Seicercus*). *Phil. Trans. R. Soc. B* **365,** 1749–1762 (2010).
8. Seddon, N. & Tobias, J. A. Character displacement from the receiver's perspective: species and mate recognition despite convergent signals in suboscine birds. *Proc. R. Soc. Lond. B* **277,** 2475–2483 (2010).
9. Grether, G. F., Losin, N., Anderson, C. N. & Okamoto, K. The role of interspecific interference competition in character displacement and the evolution of competitor recognition. *Biol. Rev. Camb. Philos. Soc.* **84,** 617–635 (2009).
10. Reznick, D. N. & Ricklefs, R. E. Darwin's bridge between microevolution and macroevolution. *Nature* **457,** 837–842 (2009).
11. Brown, W. L. & Wilson, E. O. Character displacement. *Syst. Zool.* **5,** 49–64 (1956).
12. Pfennig, K. S. & Pfennig, D. W. Character displacement: ecological and reproductive responses to a common evolutionary problem. *Q. Rev. Biol.* **84,** 253–276 (2009).
13. Gavrilets, S. & Losos, J. B. Adaptive radiation: contrasting theory with data. *Science* **323,** 732–737 (2009).
14. Dayan, T. & Simberloff, D. Ecological and community-wide character displacement: the next generation. *Ecol. Lett.* **8,** 875–894 (2005).
15. Schoener, T. W. The evolution of bill size differences among sympatric congeneric species of birds. *Evolution* **19,** 189–213 (1965).
16. Sætre, G. P. *et al.* A sexually selected character displacement in flycatchers reinforces premating isolation. *Nature* **387,** 589–592 (1997).
17. Davies, T., Meiri, S., Barraclough, T. & Gittleman, J. Species co-existence and character divergence across carnivores. *Ecol. Lett.* **10,** 146–152 (2007).
18. Connell, J. H. Diversity and the coevolution of competitors, or the ghost of competition past. *Oikos* **35,** 131–138 (1980).
19. Rundell, R. J. & Price, T. D. Adaptive radiation, non-adaptive radiation, ecological speciation and non-ecological speciation. *Trends Ecol. Evol.* **24,** 394–399 (2009).
20. Derryberry, E. P. *et al.* Lineage diversification and morphological evolution in a large-scale continental radiation: the Neotropical ovenbirds and woodcreepers (Aves: Furnariidae). *Evolution* **65,** 2973–2986 (2011).
21. Seddon, N. Ecological adaptation and species recognition drive vocal evolution in Neotropical suboscine birds. *Evolution* **59,** 200–215 (2005).
22. Tobias, J. A. *et al.* Song divergence by sensory drive in Amazonian birds. *Evolution* **64,** 2820–2839 (2010).
23. Derryberry, E. P. *et al.* Correlated evolution of beak morphology and song in the Neotropical woodcreeper radiation. *Evolution* **66,** 2784–2797 (2012).
24. Huntley, J. W. *et al.* Testing limiting similarity in Quaternary terrestrial gastropods. *Paleobiology* **34,** 378–388 (2008).
25. Monroe, M. J. Does competition drive character differences between species on a macroevolutionary scale? *J. Evol. Biol.* **25,** 2341–2347 (2012).
26. Pigot, A. L. & Tobias, J. A. Species interactions constrain geographic range expansion over evolutionary time. *Ecol. Lett.* **16,** 330–338 (2012).
27. Tobias, J. A. & Seddon, N. Signal design and perception in *Hypocnemis* antbirds: evidence for convergent evolution via social selection. *Evolution* **63,** 3168–3189 (2009).
28. Laiolo, P. Interspecific interactions drive cultural co-evolution and acoustic convergence in syntopic species. *J. Anim. Ecol.* **81,** 594–604 (2012).
29. Weir, J. T., Wheatcroft, D. & Price, T. The role of ecological constraint in driving the evolution of avian song frequency across a latitudinal gradient. *Evolution* **66,** 2773–2783 (2012).
30. Muschick, M., Indermaur, A. & Salzburger, W. Convergent evolution within an adaptive radiation of cichlid fishes. *Curr. Biol.* **22,** 2362–2368 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank G. Grether, J. Hadfield, S. Nakagawa, A. Phillimore, A. Pigot, R. Ricklefs, G. Thomas and S. West for comments and discussion. We are also indebted to the many individuals who collected specimens, tissue samples and sound recordings, and to numerous institutions (particularly the Macaulay Library, Cornell University) for granting access to this material. Complete acknowledgements and data sets are provided in the Supplementary Information. This research was supported by the John Fell Fund (to J.A.T.), the Browne Fellowship, Queen's College, Oxford, and Vetenskapsrådet (to C.K.C.), the National Science Foundation (to R.T.B.) and the Royal Society (to N.S.).

Author Contributions J.A.T. and N.S. conceived and designed the study, compiled and analysed song data, and integrated all data sets; S.C. provided morphometric data; E.P.D., S.C. and R.T.B. conducted molecular sequencing and phylogenetic analyses; C.C. designed and conducted statistical analyses, with significant input from N.S.; N.S., J.A.T. and C.C. produced figures and tables; J.A.T. prepared and edited the manuscript, with input from all authors.

Author Information Nuclear and mitochondrial DNA sequences for all lineages have been deposited in GenBank under accession numbers given in Supplementary Data 1. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.A.T. (joseph.tobias@zoo.ox.ac.uk).

LETTER

doi:10.1038/nature12977

Disease associations between honeybees and bumblebees as a threat to wild pollinators

M. A. Fürst[1,2], D. P. McMahon[3], J. L. Osborne[4,5], R. J. Paxton[3,6,7] & M. J. F. Brown[1]

Emerging infectious diseases (EIDs) pose a risk to human welfare, both directly[1] and indirectly, by affecting managed livestock and wildlife that provide valuable resources and ecosystem services, such as the pollination of crops[2]. Honeybees (*Apis mellifera*), the prevailing managed insect crop pollinator, suffer from a range of emerging and exotic high-impact pathogens[3,4], and population maintenance requires active management by beekeepers to control them. Wild pollinators such as bumblebees (*Bombus* spp.) are in global decline[5,6], one cause of which may be pathogen spillover from managed pollinators like honeybees[7,8] or commercial colonies of bumblebees[9]. Here we use a combination of infection experiments and landscape-scale field data to show that honeybee EIDs are indeed widespread infectious agents within the pollinator assemblage. The prevalence of deformed wing virus (DWV) and the exotic parasite *Nosema ceranae* in honeybees and bumblebees is linked; as honeybees have higher DWV prevalence, and sympatric bumblebees and honeybees are infected by the same DWV strains, *Apis* is the likely source of at least one major EID in wild pollinators. Lessons learned from vertebrates[10,11] highlight the need for increased pathogen control in managed bee species to maintain wild pollinators, as declines in native pollinators may be caused by interspecies pathogen transmission originating from managed pollinators.

Trading practices in domesticated animals enable infectious diseases to spread rapidly and to encounter novel hosts in newly sympatric wildlife[12]. This 'spillover' of infectious disease from domesticated livestock to wildlife populations is one of the main sources of emerging infectious disease (EID)[13]. Small or declining populations are particularly challenged, as the source host may act as a disease reservoir[14], giving rise to repeated spillover events and frequent disease outbreaks that, in the worst case, might drive already vulnerable or unmanaged populations to extinction[14]. Such severe impacts have been well documented over the past decades in vertebrates[10], but have largely been overlooked in invertebrates[15]. Recent years have seen elevated losses in multiple populations of one of the major crop-pollinating insects, the honeybee (*Apis mellifera*)[16]. EIDs have been suggested as key drivers of decline, and deformed wing virus (DWV) (particularly in combination with the exotic *Varroa* mite (*Varroa destructor*)) and *Nosema ceranae* are two likely causes for losses of honeybees[17]. As generalist pollinators, honeybees are traded and now distributed almost worldwide for crop pollination and hive products. They share their diverse foraging sites with wild pollinators and thus facilitate interspecific transmission of pathogens, as has been suggested for intraspecific disease transmission from commercial to wild bumblebee populations[18]. Our focus is on interspecific transmission, as EIDs in managed honeybees are a potential threat to a range of wild pollinators worldwide. Although evidence from small-scale studies suggests that wild pollinators like *Bombus* spp. may already harbour some honeybee pathogens[7,8,19,20], the true infectivity and landscape-scale distribution of these highly virulent EIDs in wild pollinator populations remains unknown.

To examine the potential for *Apis* pathogens to cross host–genus boundaries, we tested the infectivity of the DWV complex (which includes the very closely related, co-occurring and recombinant *Varroa destructor* virus (VDV)[21,22]; we will refer to the DWV complex as 'DWV' throughout the text) and *N. ceranae*, in controlled inoculation experiments, to one of the most common *Bombus* species in the United Kingdom (*B. terrestris*). DWV is infective for *B. terrestris*; we found significantly more DWV infections 21 days after inoculating *B. terrestris* workers versus controls (likelihood ratio test comparing the full model to one with only the intercept: $X^2 = 5.73$, d.f. $= 1$, $P < 0.017$; Fig. 1) and mean survival was reduced by 6 days. As for *Apis,* DWV causes deformed wings in *Bombus* when overtly infected[8], resulting in non-viable offspring and reduced longevity (Fig. 1). *N. ceranae* is also infective for *B. terrestris*; infections increased in *Bombus* versus control ($X^2 = 17.76$, d.f. $= 1$, $P < 0.001$; Fig. 1), although overt symptoms were not seen (mean survival increased by 4 days).

After we established that both DWV and *N. ceranae* are infective for *B. terrestris*, we conducted a structured survey across 26 sites in Great Britain and the Isle of Man (see Extended Data Fig. 1). We collected 10 *Apis* samples and 20 *Bombus* samples per site to assess EID prevalence (for details on the species identity across sites, see Extended Data Fig. 1). We analysed a total of 745 bees from 26 sites for DWV presence, DWV infection (replicating DWV) and *N. ceranae* presence. DWV was present in 20% (95% confidence interval (CI), 17–23%) of all samples; 36% (95% CI, 30–43%) of *Apis* and 11% (95% CI, 9–15%) of *Bombus*. Of the *Apis* harbouring DWV, 88% (95% CI, 70–98%) of the samples tested had actively replicating virus, whereas 38% (95% CI, 25–53%) of *Bombus* harbouring DWV had replicating virus (see Extended Data Fig. 2 and Extended Data Table 1). *N. ceranae* was less frequent, being detected in 7% (95% CI, 6–10%) of all samples; 9% (95% CI, 6–13%) of *Apis* samples and 7% (95% CI, 5–9%) of *Bombus* samples.

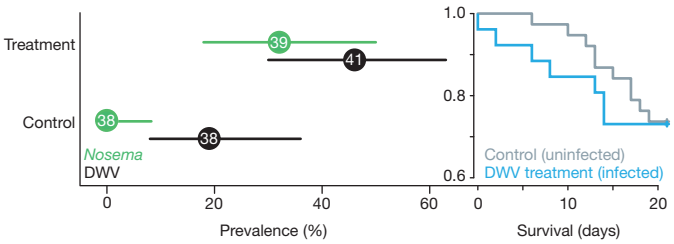


Figure 1 | DWV and *N. ceranae* infectivity in bumblebees. Prevalence of infections in treated *Bombus terrestris* workers 21 days after inoculation. Bars indicate 95% confidence intervals. Green, *Nosema*-treated samples; black, DWV-treated samples. Sample sizes are given inside the mean data point. The survival graph over the 21-day test period shows uninfected control treatments in grey compared to infected DWV treatments in blue (Cox mixed-effects model fitted with penalized partial likelihood: $X^2 = 11.93$, d.f. $= 4.17$; $P < 0.021$, see Methods) (y axis shows survival probability).

[1]Royal Holloway University of London, School of Biological Sciences, Bourne Building, Egham TW20 0EX, UK. [2]IST Austria (Institute of Science and Technology Austria), 3400 Klosterneuburg, Austria. [3]Queen's University Belfast, School of Biological Sciences, 97 Lisburn Road, Belfast BT9 7BL, UK. [4]Rothamsted Research, Department of Agro-Ecology, Harpenden AL5 2JQ, UK. [5]University of Exeter, Environment & Sustainability Institute, Penryn TR10 9EZ, UK. [6]Martin-Luther-Universität Halle-Wittenberg, Institute for Biology/General Zoology, Hoher Weg 8, 06120 Halle (Saale), Germany. [7]German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany.

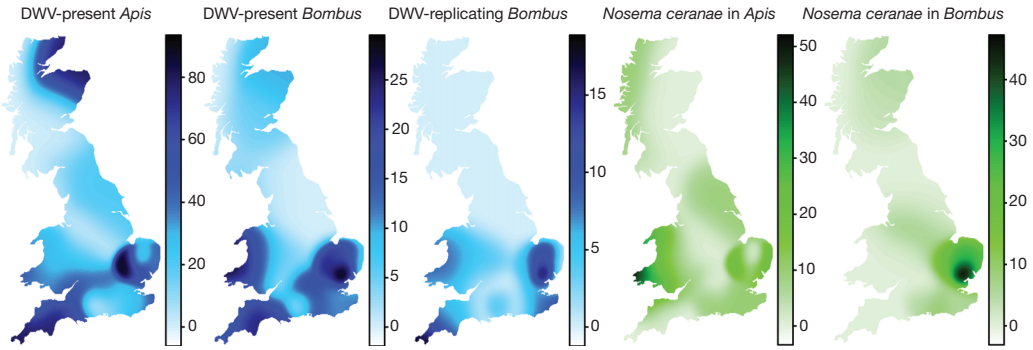


Figure 2 | Geographical distribution of DWV and *N. ceranae* across their pollinator hosts. Estimated pathogen prevalence in *Apis* and *Bombus* across Great Britain. Colour gradient (based on Gaussian kernel estimators with an adaptive bandwidth of equal number of observations over 26 sites, see Methods) corresponds to per cent prevalence (note different scales). DWV prevalence is displayed in blue and *Nosema* prevalence in green.

We estimated the Great-Britain-wide prevalence of the two pathogens in *Apis* and *Bombus* spp. based on our field survey data (Fig. 2). We found no evidence for spatial clustering of DWV presence in *Bombus* (Moran's $I = 0.023$, $P > 0.211$) or either of the pathogens in *Apis* (DWV presence: Moran's $I = 0.03$, $P > 0.186$; *Nosema* presence: Moran's $I = -0.061$, $P > 0.649$). However, there was weak clustering of DWV infection in *Bombus* (Moran's $I = 0.061$, $P < 0.044$) and very strong clustering of *N. ceranae* in *Bombus* (Moran's $I = 0.25$, $P < 0.001$), indicating disease hotspots for DWV in *Bombus* in the south west and east of Great Britain and for *N. ceranae* in *Bombus* in the south east of Great Britain (Fig. 2). As prevalence was lower in *Bombus* than *Apis*, we modelled pathogen prevalence in *Bombus* as dependent on pathogen prevalence in *Apis, Apis*:*Bombus* (ratio of host densities) and *Apis* abundance, including biologically relevant interactions while controlling for latitude, longitude and sunlight hours, and adding collection site and species identity as random factors. Our full model for DWV presence fitted the data significantly better than the null model without any of the test predictors and their interactions included (likelihood ratio test: $X^2 = 19.03$, d.f. $= 5$, $P < 0.002$). After removal of the non-significant interactions (general linear mixed model (GLMM): *Bombus*:*Apis* × DWV presence in *Apis*, estimate ± s.e. of the estimate of the fixed effect parameter in the model $= -0.105 \pm 1.376$, $P = 0.939$; *Apis* abundance × DWV presence in *Apis*, 0.425 ± 1.309, $P = 0.745$), it is clear that prevalence of DWV in *Apis* has a strong positive effect on DWV prevalence in *Bombus* (GLMM: 2.718 ± 0.921, $z = 2.951$, $P < 0.004$) (Fig. 2, and Extended Data Fig. 3), whereas none of the other predictors had a role (GLMM: *Bombus*:*Apis*, 0.315 ± 0.387, $z = 0.814$, $P < 0.416$; *Apis* abundance, -0.085 ± 0.364, $z = -0.232$, $P < 0.816$). In the case of *N. ceranae*, our full model fitted the data significantly better than the null model ($X^2 = 15.8$, d.f. $= 5$, $P < 0.008$). Specifically, there was an effect of *Nosema* prevalence in *Apis* on *Nosema* prevalence in *Bombus* and this varied with *Apis* abundance (interaction between *Nosema* prevalence in *Apis* and *Apis* abundance, $X^2 = 7.835$, d.f. $= 2$, $P < 0.02$), whereas *Bombus*:*Apis* did not explain *Nosema* prevalence in *Bombus* (GLMM: 8.386 ± 6.793, $z = 1.235$, $P = 0.217$) (Fig. 2 and Extended Data Fig. 3).

The prevalence data implied local transmission of DWV between *Apis* and *Bombus*. To test this, we sequenced up to five isolates per DWV infected *Bombus* sample from five sites matched by up to five isolates of sympatric DWV infected *Apis* samples. If a pathogen is transmitted between these two hosts, we would expect *Apis* and *Bombus* to share the same DWV strain variants within a site. Marginal log likelihoods estimated by stepping stone sampling[23] decisively supported clades constrained by site as opposed to host, indicating pathogen transmission within site (Fig. 3 and Extended Data Table 2).

Our results provide evidence for an emerging pathogen problem in wild pollinators that may be driven by *Apis*. Our data cannot demonstrate directionality in the interspecific transmission of DWV. However, the high prevalence of DWV in honeybees, which is a consequence of the exotic vector *Varroa destructor*[24], is consistent with the hypothesis that they are the major source of infection for the pollinator community. Similar results have been found for intraspecific transmission of *Bombus*-specific pathogens from high prevalence commercial *Bombus* colonies to low prevalence wild *Bombus* populations[18]. Our field estimates

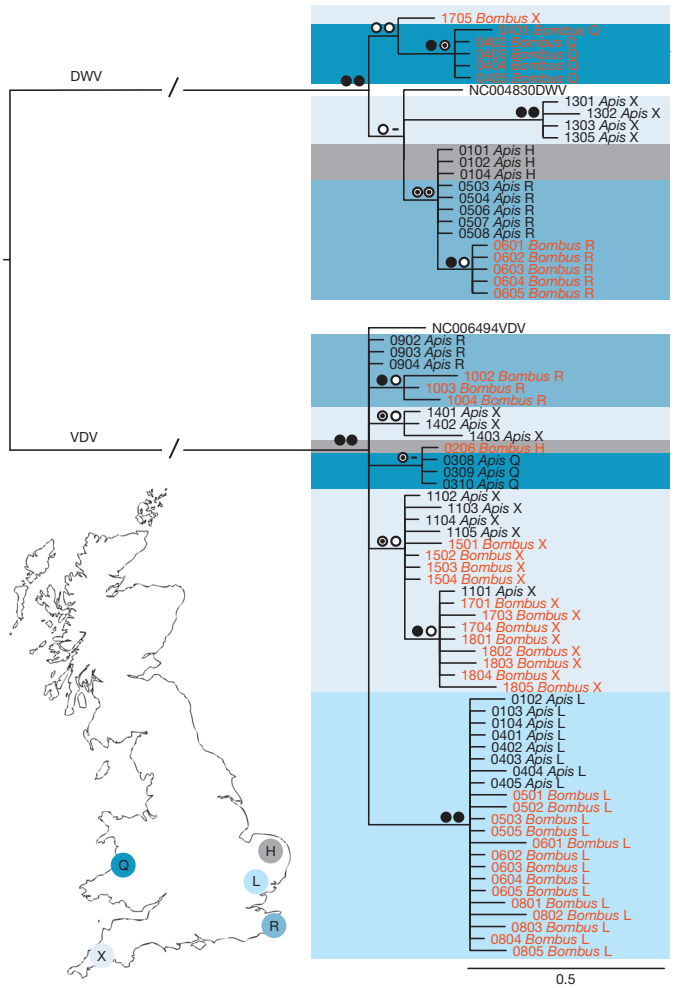


Figure 3 | Sympatric *Apis* and *Bombus* share viral strains. RNA-dependent RNA polymerase partial gene phylogeny of pollinator viruses (see main text). Gene trees were estimated using PhyML v.3.0 maximum-likelihood (ML) bootstrapping (500 replicates) and MrBayes v3.2.1 (see Methods). Coloured boxes correspond to sites H, L, Q, R and X (as shown on the map) and text colours correspond to host (red, *Bombus*; black, *Apis*). Symbols represent node support values: posterior probability (left), bootstrap support (right). Filled circle, >90%; target symbol, >70%; empty circle, >50%; hyphen, <50%. Branches (/), one-third of true length. Numbers are IDs; first two numbers represent the individual, second two numbers represent the clone.

of prevalence are conservative for DWV, as highly infected individuals have deformed wings, are incapable of flight, and thus would not be captured by our sampling protocol. Consequently, DWV prevalence and, as a result, impact are likely to be higher in managed and wild populations than our data suggest. Interestingly, *N. ceranae* prevalence in *Bombus* depends positively on *Apis* abundance, but only when *N. ceranae* prevalence in *Apis* is low, suggesting a possible environmental saturation effect of *N. ceranae* spores. In contrast to the low impact of *N. ceranae* on the survival of *B. terrestris* in our study, very high virulence was found by another study[25]. This might be explained by our use of young bees compared to the non-age-controlled design used in the other paper[25], indicating age-dependent differential susceptibility in *B. terrestris,* as has been suggested to be the case in honeybees[26].

Ongoing spillover of EIDs could represent a major cause of mortality of wild pollinators wherever managed bees are maintained. Although our data are only drawn from Great Britain, the prerequisites for honeybees to be a source or reservoir for these EIDs—high colony densities and high parasite loads—are present at a global scale. In addition, global trade in both honeybees and commercial *Bombus* may exacerbate this impact[6,27]. Reducing the pathogen burden in managed honeybees so as to reduce the risk of transmission to wild pollinators is not straightforward. Tighter control of importation and hygiene levels of transported colonies could be imposed with regulation, but policies developed in this direction must learn from the past; such regulation is difficult to implement and hard to evaluate[9,28]. Clearly, it is essential to ensure that those managing bees (including commercial producers, growers and beekeepers) have access to the methods and skills to monitor, manage and control EIDs for the benefit of their managed colonies, and the wider pollinator community. A consensus on the threat of EIDs for wild pollinators can only be reached with greater knowledge of their epidemiology, global extent and impact, and it will be crucial to involve key stakeholders (for example, the beekeeping community, *Bombus* exporters) in any decision process, as any progress made will largely be driven by their actions.

METHODS SUMMARY

***Bombus* inoculation experiment.** Two-day-old workers of *Bombus terrestris audax* colonies (Biobest) were individually inoculated with either 10^5 spores per bee purified *N. ceranae* or 10^9 genome equivalents per bee purified DWV in 10 μl sucrose solution. Bees surviving for 21 days were freeze-killed and tested for pathogen presence using molecular techniques.

Sampling scheme. Sampling took place at 24 mainland sites and 2 islands, Colonsay and the Isle of Man, which are currently free of *Varroa destructor* (the main vector for DWV in *Apis mellifera*) (see Extended Data Fig. 1 for site distribution). Cryptic *Bombus* species were identified by PCR–RFLP (restriction fragment length polymorphism) analysis. *Apis* and *Bombus* densities were estimated for each site by timing the collection effort for 20 samples from each genus simultaneously. Samples collected were freeze-killed at −20 °C and transferred to −80 °C as soon as possible thereafter. RNA and DNA preparation and virus strand specific PCR with reverse transcription (RT–PCR) followed standard protocols.

Statistics. True prevalences with 95% confidence intervals were computed with the function epi.prev in the R library epiR, version 0.9-45.

Overall prevalence for each of our parasites was calculated using Gaussian kernel estimators with an adaptive bandwidth of equal number of observations (set to 3× the maximum observations per site) (R library prevR, version 2.1, function kde).

Moran's *I* was calculated as implemented in the R library package ape (version 3.0-7, function Moran.I).

We ran GLMMs to investigate both effects on disease status of individuals 21 days after pathogen challenge and also pathogen prevalence in *Bombus* using the function lmer of the R package lme4. All analyses were run in R.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 October; accepted 30 December 2013.

1. Binder, S., Levitt, A. M., Sacks, J. J. & Hughes, J. M. Emerging infectious diseases: public health issues for the 21st century. *Science* **284,** 1311–1313 (1999).
2. Oldroyd, B. P. Coevolution while you wait: *Varroa jacobsoni*, a new parasite of western honeybees. *Trends Ecol. Evol.* **14,** 312–315 (1999).
3. Ratnieks, F. L. W. & Carreck, N. L. Clarity on honey bee collapse? *Science* **327,** 152–153 (2010).
4. Vanbergen, A. J. & The Insect Pollinator Initiative. Threats to an ecosystem service: pressures on pollinators. *Front. Ecol. Environ* **11,** 251–259 (2013).
5. Williams, P. H. & Osborne, J. L. Bumblebee vulnerability and conservation world-wide. *Apidologie* **40,** 367–387 (2009).
6. Cameron, S. A. *et al.* Patterns of widespread decline in North American bumble bees. *Proc. Natl Acad. Sci. USA* **108,** 662–667 (2011).
7. Evison, S. E. F. *et al.* Pervasiveness of parasites in pollinators. *PLoS ONE* **7,** e30641 (2012).
8. Genersch, E., Yue, C., Fries, I. & de Miranda, J. R. Detection of deformed wing virus, a honey bee viral pathogen, in bumble bees (*Bombus terrestris* and *Bombus pascuorum*) with wing deformities. *J. Invertebr. Pathol.* **91,** 61–63 (2006).
9. Meeus, I., Brown, M. J. F., De Graaf, D. C. & Smagghe, G. Effects of invasive parasites on bumble bee declines. *Conserv. Biol.* **25,** 662–671 (2011).
10. Fisher, M. C. *et al.* Emerging fungal threats to animal, plant and ecosystem health. *Nature* **484,** 186–194 (2012).
11. Krebs, J. *et al. Bovine Tuberculosis in Cattle and Badgers* (MAFF Publications, 1997).
12. Vitousek, P. M., Dantonio, C. M., Loope, L. L. & Westbrooks, R. Biological invasions as global environmental change. *Am. Sci.* **84,** 468–478 (1996).
13. Daszak, P. Emerging infectious diseases of wildlife — threats to biodiversity and human health. *Science* **287,** 443–449 (2000).
14. Dobson, A. Population dynamics of pathogens with multiple host species. *Am. Nat.* **164,** S64–S78 (2004).
15. Alderman, D. J. Geographical spread of bacterial and fungal diseases of crustaceans. *Rev. Sci. Tech.* **15,** 603–632 (1996).
16. Neumann, P. & Carreck, N. L. Honey bee colony losses. *J. Apic. Res.* **49,** 1–6 (2010).
17. Paxton, R. J. Does infection by *Nosema ceranae* cause "Colony Collapse Disorder" in honey bees (*Apis mellifera*)? *J. Apic. Res.* **49,** 80–84 (2010).
18. Murray, T. E., Coffey, M. F., Kehoe, E. & Horgan, F. G. Pathogen prevalence in commercially reared bumble bees and evidence of spillover in conspecific populations. *Biol. Conserv.* **159,** 269–276 (2013).
19. Singh, R. *et al.* RNA viruses in Hymenopteran pollinators: evidence of inter-taxa virus tansmission via pollen and potential impact on non-*Apis* Hymenopteran species. *PLoS ONE* **5,** e14357 (2010).
20. Graystock, P. *et al.* The Trojan hives: pollinator pathogens, imported and distributed in bumblebee colonies. *J. Appl. Ecol.* **50,** 1207–1215 (2013).
21. Ongus, J. R. *et al.* Complete sequence of a picorna-like virus of the genus Iflavirus replicating in the mite *Varroa destructor. J. Gen. Virol.* **85,** 3747–3755 (2004).
22. Moore, J. *et al.* Recombinants between deformed wing virus and *Varroa destructor* virus-1 may prevail in *Varroa destructor*-infested honeybee colonies. *J. Gen. Virol.* **92,** 156–161 (2011).
23. Xie, W., Lewis, P. O., Fan, Y., Kuo, L. & Chen, M.-H. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* **60,** 150–160 (2011).
24. Martin, S. J. *et al.* Global honey bee viral landscape altered by a parasitic mite. *Science* **336,** 1304–1306 (2012).
25. Graystock, P., Yates, K., Darvill, B., Goulson, D. & Hughes, W. O. H. Emerging dangers: deadly effects of an emergent parasite in a new pollinator host. *J. Invertebr. Pathol.* **114,** 114–119 (2013).
26. Smart, M. D. & Sheppard, W. S. *Nosema ceranae* in age cohorts of the western honey bee (*Apis mellifera*). *J. Invertebr. Pathol.* **109,** 148–151 (2012).
27. Otterstatter, M. C. & Thomson, J. D. Does pathogen spillover from commercially reared bumble bees threaten wild pollinators? *PLoS ONE* **3,** (2008).
28. Donnelly, C. A. & Woodroffe, R. Reduce uncertainty in UK badger culling. *Nature* **485,** 582 (2012).

Acknowledgements The authors are grateful to E. Fürst for technical support and R. J. Gill for discussions. We thank C. Jones, G. Baron and O. Ramos-Rodriguez for comments on previous versions of the manuscript. They also thank Hymettus Ltd for help with the field collections, K. Liu for help in the laboratory and B. McCrea and S. Baldwin for technical help in the bee laboratory. The study was supported by the Insect Pollinators Initiative (funded jointly by the Biotechnology and Biological Sciences Research Council, the Department for Environment, Food and Rural Affairs, the Natural Environment Research Council, The Scottish Government and The Wellcome Trust, under the Living with Environmental Change Partnership: grants BB/I000151/1 (M.J.F.B.), BB/I000100/1 (R.J.P.) and BB/I000097/1 (J.L.O.).

Author Contributions The study was jointly conceived by R.J.P., J.L.O. and M.J.F.B. Experiments were designed by M.A.F. and M.J.F.B.; M.A.F prepared the manuscript; M.J.F.B., D.P.M., R.J.P. and J.L.O. edited the manuscript. M.A.F. carried out the experimental, molecular work and analyses, and D.P.M. undertook the phylogenetic analyses.

Author Information Viral RNA sequences have been deposited in GenBank under accession numbers KF929216–KF929290. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.A.F. (Apocrite@gmail.com).

METHODS

***Bombus* inoculation experiment.** Each of the seven experimental *Bombus terrestris* colonies (Biobest) was tested for the presence of the two treatment pathogens DWV and *N. ceranae*. Daily, callows (newly emerged workers) were removed from the colony, assigned sequentially to random treatment blocks and housed individually in small Perspex boxes on an *ad libitum* diet of 50% sucrose solution and artificial pollen (Nektapoll), as natural pollen has been shown to contain viable *N. ceranae* spores and DWV virions[19,29]. Two-day-old bumblebee workers were individually inoculated with a treatment-dependent inoculum in 10 µl sucrose. Crude hindgut extracts of five *Apis* workers propagating *N. ceranae* were purified by the triangulation method[30] with slight adaptations.

We used small cages with 30 *N. ceranae* infected honeybees to propagate *N. ceranae* spores for the inoculum. Every second day we collected five honeybees from these cages, and removed and ground the hindguts. The resulting extract was filtered through cotton and washed with 0.9% insect ringer (Sigma Aldrich). We triangulated extracts using Eppendorf tubes and spin speeds of 0.5 RCF (relative centrifugal force) for 3 min, purifying *N. ceranae* spores across a series of seven tubes. Spore numbers were quantified in a Neubauer counting chamber. In parallel, we extracted and purified *N. ceranae* free bees to use for control inoculations.

DWV virus inoculum was prepared according to a previous paper[31] but with modifications. Honeybees with DWV symptoms (crippled wings and body deformities) were crushed in 0.5 M potassium phosphate buffer, pH 8.0, filtered and clarified by slow-speed centrifugation (8,000*g* for 10 min) before being diluted and injected (1 µl) into white-eyed pupae for bulk propagation of virus. After 5 days, up to 100 pupae were harvested, and after a screen by quantitative RT–PCR (qRT–PCR), virus was purified. Virus extraction buffer consisted of 0.5 M potassium phosphate pH 8.0, 0.2% DEICA, 10% diethyl ether. Purification consisted of two slow-speed clarifications (8,000*g* for 10 min), one high-speed clarification (75,000*g* for 3 h) followed by re-suspension in 0.5 M potassium phosphate buffer (pH 8.0) and a final slow speed clarification. Virus preparations were aliquoted and stored at −80 °C until use in inoculation experiments.

Quantitative RT-PCR was used to check the purified virus for presence of DWV and absence of other common honeybee RNA viruses: ABPV (acute bee paralysis virus), BQCV (black queen cell virus), CBPV (chronic bee paralysis virus), IAPV (Israeli acute paralysis virus), SBV (sacbrood virus) and SBPV (slow bee paralysis virus).

A duplicate dilution series of external DNA standards covering 10^2 to 10^8 molecules (reaction efficiencies: 90–110%, r^2: 0.95–0.99) were included in qRT–PCR runs to quantify DWV genome equivalents present in the inoculum. For absolute quantification of virus dose, an external DNA standard was generated by amplifying a genomic fragment of 241 bp using the primers F8668*std* (5′-GATGGG TTTGATTCGATATCTTGG-3′) and B8757*std* (5′-GGCAAACAAGTATCTTT CAAACAATC-3′) via RT–PCR that contained the 136-bp fragment amplified by the DWV-specific qRT–PCR primers F8668 and B8757 (ref. 32).

Shortly before administration, inocula were prepared to a total concentration of 10^5 spores per bee in 10 µl (10^4 spores per µl sucrose solution). Inocula were administered individually in a small Petri dish after 30–60 min starvation. Only workers ingesting the full 10 µl within 1 h were used in the experiment.

Sampling scheme. The mainland sampling sites were chosen across Great Britain along a north–south transect (12 sampling points with fixed latitude, but free in longitude) and across two east–west transects (12 sampling points with fixed longitude, but free within a narrow latitudinal corridor). Each of the mainland sites were at least 30 km apart (mean ± s.d. of nearest neighbour = 69.21 ± 26.39). The island sites were chosen deliberately to gain background data for both *Apis* and *Bombus* disease prevalence in the absence of *Varroa*, the main transmission route for DWV in *Apis*. At each sampling site we collected approximately 30 workers for each of the following species: *Apis mellifera, Bombus terrestris* (verified by RFLP analysis[33]), and the next most common bumblebee on site. We collected free-flying bees from flowers rather than bees from colonies as this is the most likely point of contact in the field. By collecting from flowers we lowered the likelihood of collecting bumblebees from different colonies. Although we ran the risk of collecting multiple honeybees from the same hive, this nevertheless represents the potential force of infection for both genera in the field.

Each collection took place along a continuous transect, where maximally 10 bees per 10-m stretch were collected before moving on to the next ten metre stretch. At each site, the collection area covered at least 1000 m^2 (for example, 10 × 100 m, 20 × 50 m). Each sampling point was within one of the following landcover types: urban areas (gardens and parks), farmland (hedgerows, border strips, crops and wildflower meadows), coastal cliffs, sand dunes and heather moorland.

If possible, we collected all bees within a single day. In the case of adverse weather, we returned as soon as possible to finish the collection at the exact same site. To estimate *Apis* and *Bombus* densities at each site we timed the collection effort simultaneously. Time taken to collect 20 *Bombus* workers (of any *Bombus* species) and 20 *Apis* workers was recorded, respectively. Timed collecting efforts took place on a single day only.

Samples collected were put in sampling tubes, transferred straight onto ice, then freeze-killed at −20°C and transferred to −80°C as soon as possible thereafter to ensure optimal RNA (DWV) preservation.

RNA work. RNA extraction followed the standard RNeasy plant mini kit (Qiagen) protocol with the final elutate (in RNase free ddH2O) of 30 µl being run over the column twice (for optimal RNA concentration). For reverse transcription of RNA to complementary DNA we followed the standard protocol of the Nanoscript Kit (Primerdesign). Our priming was target specific in separate reactions for *N. ceranae* (primer pair *N. ceranae*[34]), DWV (primer pair F15–B23 (ref. 35)) and a housekeeping gene (primer pair ACTB[36]) as a positive control for RNA extraction efficiency. Bees were transferred to liquid N_2 before dissection. Each bee's abdomen was cut with a sterile scalpel dorsoventrally along the sagittal plane. One half was submerged in RLT buffer (Qiagen) for RNA extraction, and the second half was archived at −80 °C. Tissue disruption and homogenization of individual half-abdomens was performed on a tissue lyser II (Qiagen) at 30 Hz for 2 min followed by 20 Hz for 2 min. RNA quality and quantity were checked on a Spectrometer (Nanodrop, Thermo Scientific). cDNA preparation was conducted at 65 °C for 5 min for the initial priming immediately before the addition of the reverse transcriptase. For the extension, samples were incubated at 25 °C for 5 min followed by 55 °C for 20 min and then heat inactivated for 15 min at 75 °C. cDNA was used as template in a standard PCR with 57 °C, 54 °C and 57 °C annealing temperatures, respectively. Results were visualized on a 2% agarose gel with ethidium bromide under ultraviolet light. Agarose gels were scored without knowledge of sample ID. To verify the specificity of the amplicon, one purified PCR product taken from *Apis* and one taken from *B. lapidarius* were sequenced (Macrogen).

Detection of negative-strand DWV. Detection of pathogens in pollinators in the field does not provide proof of infection, as pathogens are likely to be ingested on shared, contaminated food resources and therefore are inevitably present in the gut lumen as passive contaminants without necessarily infecting the host. To minimize these cases, we tested all of our DWV positive *Bombus* samples and a subset of DWV positive *Apis* samples for virus replication, a strong indicator for infection[37]. DWV is a positive strand virus whose negative strand is only present in a host once the virus is actively replicating[31]. Reverse transcription was conducted using a tagged primer tagB23 (ref. 38) for the initial priming to target exclusively the negative strand. The resulting cDNA was used in a PCR with the tag sequence and F15 as primers[38,39]. We tested all *Bombus* samples that were positive for DWV presence and, where possible, two DWV-positive *Apis* samples from each site where we found DWV in *Bombus*.

Sequencing. DWV sequence diversity was analysed by sequencing up to five independent clones per *Bombus* samples infected with negative-strand DWV from five sites (H, L, Q, R, X; chosen for their high DWV-infection prevalence in *Bombus*) and five clones of DWV-infected *Apis* samples from the same sites (we checked extra *Apis* samples for DWV infection, if necessary, to match *Bombus* DWV infections). All *Bombus* samples were *B. lapidarius*, with the exception of one sample from site L (clone 05), which was *B. pascuorum* (this sample is not included in any of the other analyses, but revealed a DWV infection in an initial screening and was hence included in the virus variant analysis). We sequenced a region of the DWV genome: the RNA-dependent RNA polymerase gene (F15–B23 primer pair[35] used throughout the study).

RNA-dependent RNA polymerase is thought to be a conserved region of the virus genome in which non-synonymous substitutions may have significant implications for the epidemiology of the virus[24]. RT–PCRs and PCR were run as described before. DWV PCR products were verified by gel electrophoresis as described above; if a clear, clean single band was visible, we proceeded directly to the cloning protocol. If not, we purified products from the agarose gel following a standard protocol (Qiaquick Gel Extraction Kit, Qiagen) and used the purified fragment in an additional PCR. PCR products were cloned using the Invitrogen TA cloning kit (Invitrogen), according to the manufacturer's instructions. Plasmid DNA was isolated using the Spin Miniprep kit (Qiagen) and the successful insertion of target sequence was tested by restriction analysis (digested with EcoRI). Up to five clones per sample were sequenced in forward and reverse orientation (Source BioSciences).

Analysis of DWV sequences. The 75 *Apis* and *Bombus* clones from sites H, L, Q, R and X were supplemented with DWV and VDV reference RNA-dependent RNA polymerase sequences (accession numbers NC004830 and NC006494, respectively), resulting in a final alignment of 420 bp from 77 sequences. Forward and reverse sequences of each clone were assembled and the consensus sequence was used for further analysis. Sequences were aligned using Geneious (R 6.1.6) with standard settings. Ends were trimmed by hand. For the tree building we conducted two independent (MC)^3 algorithms running for two-million generations, each with four chains (three hot, one cold), sampling 1 tree in 1,000, under the GTR + Γ (generalised time reversible model of sequence evolution with a gamma distribution)

(Nst (number of substitution types) = 6) substitution model. Gene trees were estimated using PhyML v.3.0 (ref. 40) maximum-likelihood (ML) bootstrapping (500 replicates) and MrBayes v3.2.1 (ref. 41), under a GTR + Γ model, using four categories to accommodate rate variation across sites. Burn-in cutoffs (the time given for the tree sampling to converge to its stationary phase, to determine the trees removed prior to analysis) were inspected manually for each parameter file in Tracer v1.5 (ref. 42). Inspection of the standard deviation of split frequencies confirmed that the trees sampled in both $(MC)^3$ runs had converged (0.0093). To test alternative a priori hypotheses of virus diversification, for each virus (DWV and VDV) we constrained clades according to site (H, L, Q, R and X) or host genus (*Apis* and *Bombus*), and performed stepping stone sampling[23] as implemented in MrBayes v3.2.1 to estimate marginal log likelihoods. Sampling was conducted for 50 steps of 39,000 generations for two independent MCMC runs to ensure that accurate estimates were obtained. The first 9,000 generations of every step were discarded as burn-in. The model with the highest likelihood score was used as the null hypothesis. We compared Bayes factors for both models and used a threshold of 2 ln (Bayes factors) > 10 as decisive support for the null against the alternative hypothesis[43] (Extended Data Table 2). We repeated stepping-stone sampling to confirm run stability (data not shown).

Statistics. Mean survival of control treatments, free of the two test pathogens, was 14.2 ± 4.2 days (mean ± s.d.), whereas DWV-treated bees survived for 8.1 ± 5.8 days (mean ± s.d.). To assess the effect of infection on survival we fitted a Cox mixed effects model with treatment as a fixed factor and colony of origin as a random factor and compared it to the null model[44] (R library coxme, version 2.2-3, function coxme). The model was fitted with the penalized partial likelihood (PPL) and showed a significant negative impact of infection on longevity ($X^2 = 11.93$, d.f. = 4.17; $P < 0.021$).

N.-ceranae-treated bees survived for 18 ± 1 days (mean ± s.d.). A model with treatment as a fixed factor and colony of origin as a random factor showed no improvement over the null model (PPL: $X^2 = 0.12$, d.f. = 1; $P > 0.735$).

True prevalences with 95% confidence intervals were computed to correct for varying sample sizes (owing to the different species of bumblebee at the sampling sites) and test sensitivity was set to a conservative 95% (ref. 45). Confidence-interval estimates are based on a previous method for exact two-sided confidence intervals[46] for each sampling site and for each species sampled[47] (R library epiR, version 0.9-45, function epi.prev).

To investigate our spatially distributed data set we undertook an exploratory data analysis (EDA)[48] in which we calculated a prevalence surface for each of our parasites using Gaussian kernel estimators with an adaptive bandwidth of equal number of observations. This is a variant of the nearest neighbour technique, with bandwidth size being determined by a minimum number of observations in the neighbourhood (set to three times the maximum observations per site)[49] (R library prevR, version 2.1, function kde). Estimated surfaces were used for visual inspection only (Fig. 2); all the remaining analyses are based on the raw data only.

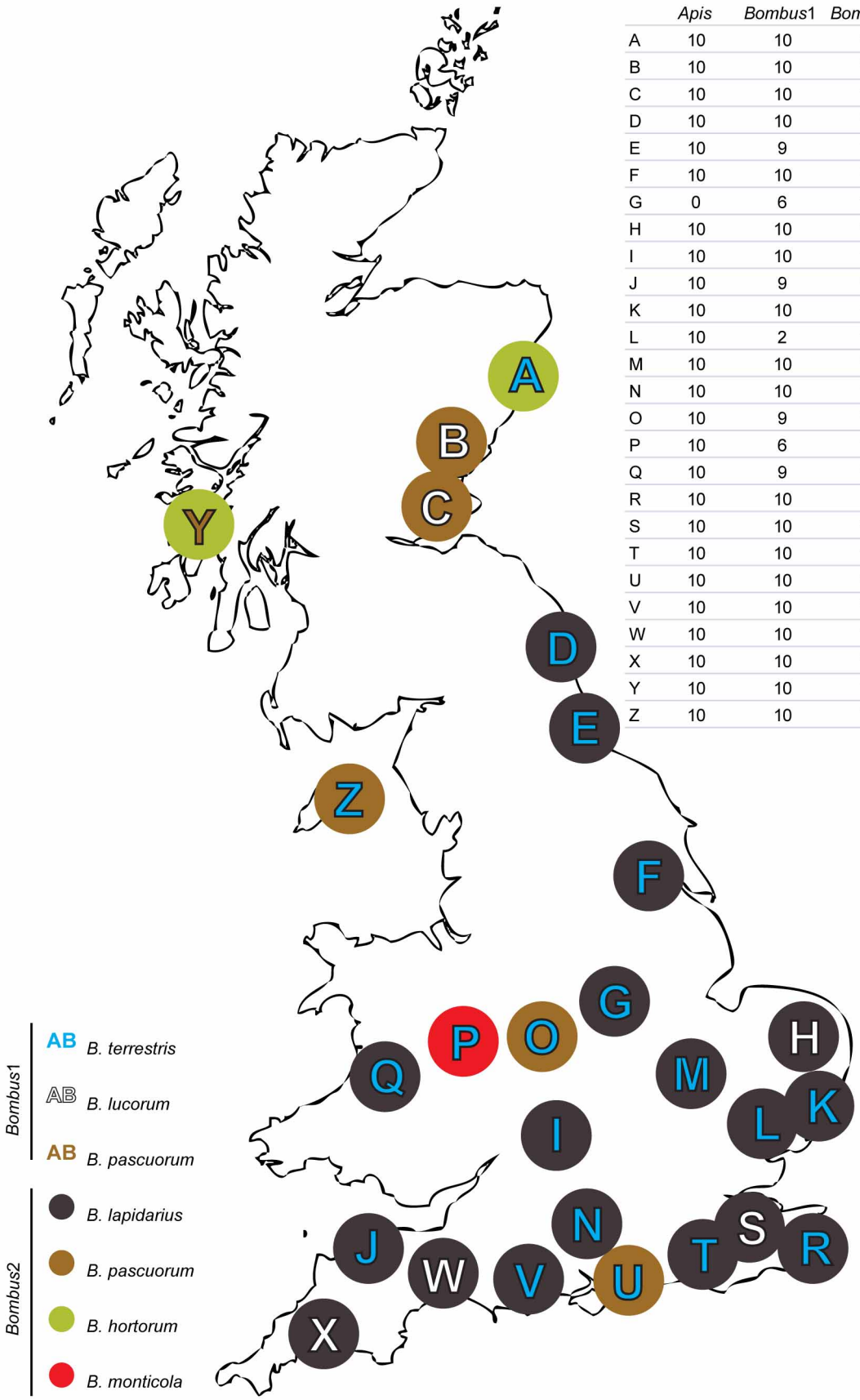
To investigate spatial structure and disease hotspots we used spatial autocorrelation statistics of the true prevalence of each of the pathogens in the different host genera from the 26 collection sites. To identify whether or not the pathogens we found were spatially clustered, we computed the spatial autocorrelation coefficient Moran's I[50] with an inverse spatial distance weights matrix, as implemented in Gittleman and Kot[51] (R library ape, version 3.0-7, function Moran.I)[52]. Moran's I is a weighted measure describing the relationship of the prevalence values associated with spatial points. The coefficient ranges from −1 (perfect dispersion) through 0 (no spatial autocorrelation (random distribution)) to 1 (perfect clustering).

To investigate whether pathogen prevalence (*Nosema* and DWV were tested in separate models) in *Apis*, *Bombus*:*Apis* (ratio of host densities), or *Apis* absolute abundance had an effect on pathogen prevalence in *Bombus*, we ran a generalized linear mixed model (GLMM)[53] with binomial error structure and logit link function using the function lmer of the R package lme4 (ref. 54). Latitude, longitude, sunlight hours (a proxy for favourable foraging weather that would enable disease transmission; calculated cumulatively from March until the month of collection (data were collected from the MET office webpage: http://www.metoffice.gov.uk/climate/uk/anomacts/, averaging over area sunlight-hour-ranges)) and landcover type were included in the model as fixed control effects (present in the full as well as the null model), and site and species were included in the model as random effects (present in the full as well as the null model). Before running the model we inspected all predictors for their distribution, as a consequence of which we log transformed '*Bombus*:*Apis*' and '*Apis* abundance' to provide more symmetrical distributions. Thereafter, we z-transformed all quantitative predictors to a mean of zero and a standard deviation of one to derive more comparable estimates and to aid interpretation of interactions[55]. As changes in '*Bombus*:*Apis*' and '*Apis* abundance' could lead to changes in pathogen prevalence in *Bombus* because of a change in pathogen prevalence in *Apis*, we included the interactions between '*Bombus*:*Apis*' and pathogen prevalence in *Apis*, and '*Apis* abundance' and pathogen prevalence in *Apis*. To test the overall effect of our three test predictors, we compared the full model with a reduced model (null model) using a likelihood ratio test comprising latitude, longitude, sunlight hours and landcover type with the same random effects structure. Model stability was assessed by excluding data points one by one and comparing the estimates derived from these reduced models with estimates from the full model (revealing a stable model). Site G had to be excluded from this analysis as no *Apis* samples were found on site.

We fitted linear models to assess the relationships of parasite prevalence among *Apis* and *Bombus*.

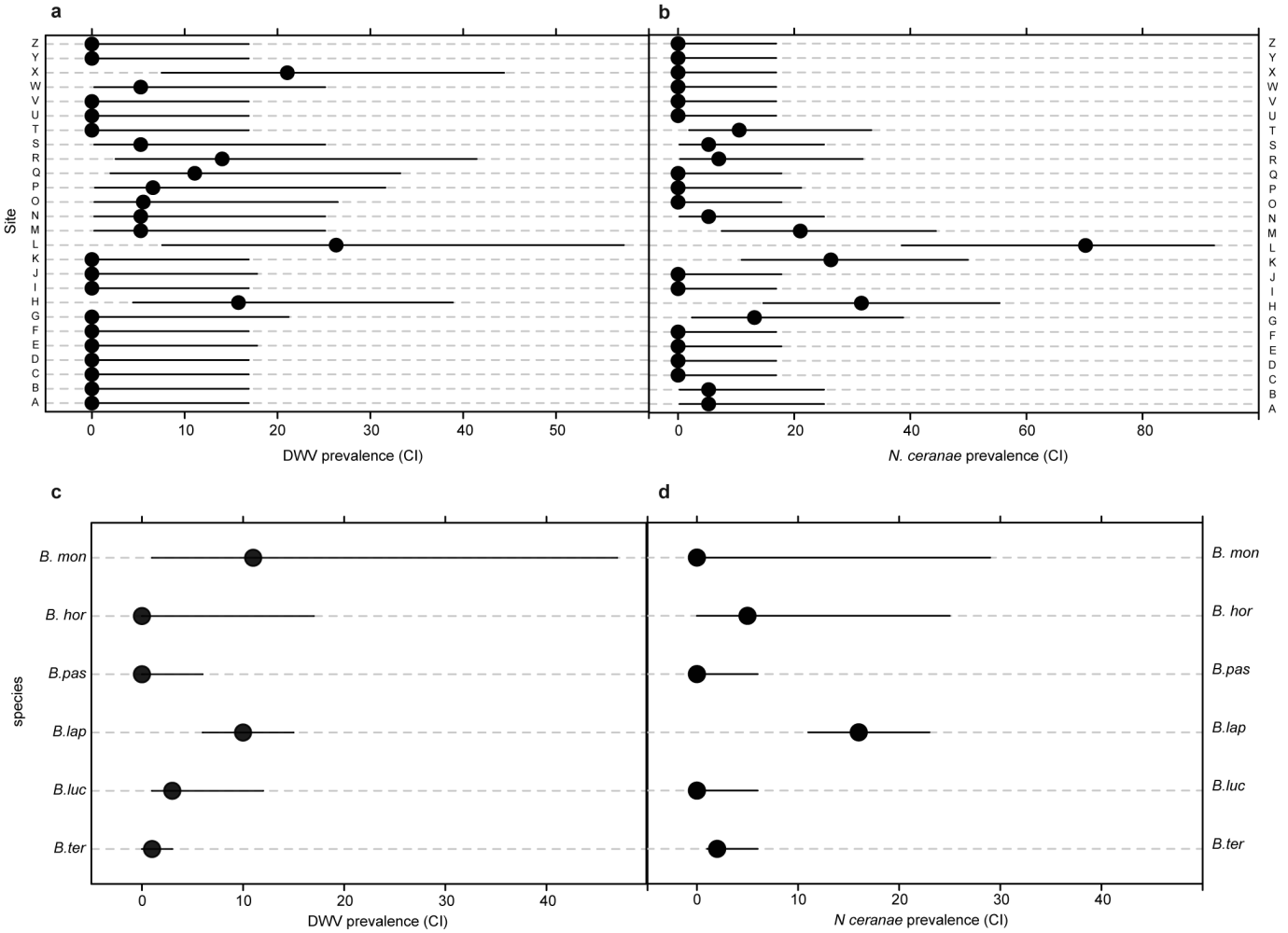
We investigated the effect of pathogen treatment on disease status of an individual with a GLMM[53] with binomial error structure and logit link function using the function lmer of the R package lme4 (ref. 54). Colony of origin was entered into the model as a random effect. As described before, we checked model stability (the model with interaction terms included was unstable; however it stabilized once the non-significant interaction terms were removed), before testing the full model against the null model using a likelihood-ratio test. All analyses were run in R (ref. 56).

29. Higes, M., Martin-Hernandez, R., Garrido-Bailon, E., Garcia-Palencia, P. & Meana, A. Detection of infective *Nosema ceranae* (Microsporidia) spores in corbicular pollen of forager honeybees. *J. Invertebr. Pathol.* **97,** 76–78 (2008).
30. Cole, R. J. Application of the ''triangulation'' method to the purification of *Nosema* spores from insect tissues. *J. Invertebr. Pathol.* **15,** 193–195 (1970).
31. Bailey, L. L. & Ball, B. V. *Honey bee pathology* 2nd edn, (Academic Press, 1991).
32. Yañez, O. *et al.* Deformed wing virus and drone mating flights in the honey bee (*Apis mellifera*): implications for sexual transmission of a major honey bee virus. *Apidologie* **43,** 17–30 (2012).
33. Murray, T. E., Fitzpatrick, U., Brown, M. J. F. & Paxton, R. J. Cryptic species diversity in a widespread bumble bee complex revealed using mitochondrial DNA RFLPs. *Conserv. Genet.* **9,** 653–666 (2008).
34. Chen, Y., Evans, J. D., Smith, I. B. & Pettis, J. S. *Nosema ceranae* is a long-present and wide-spread microsporidian infection of the European honey bee (*Apis mellifera*) in the United States. *J. Invertebr. Pathol.* **97,** 186–188 (2008).
35. Genersch, E. Development of a rapid and sensitive RT-PCR method for the detection of deformed wing virus, a pathogen of the honeybee (*Apis mellifera*). *Vet. J.* **169,** 121–123 (2005).
36. Hornáková, D., Matouskova, P., Kindl, J., Valterova, I. & Pichova, I. Selection of reference genes for real-time polymerase chain reaction analysis in tissues from *Bombus terrestris* and *Bombus lucorum* of different ages. *Anal. Biochem.* **397,** 118–120 (2010).
37. de Miranda, J. R. & Genersch, E. Deformed wing virus. *J. Invertebr. Pathol.* **103,** S48–S61 (2010).
38. Yue, C. & Genersch, E. RT-PCR analysis of deformed wing virus in honeybees (*Apis mellifera*) and mites (*Varroa destructor*). *J. Gen. Virol.* **86,** 3419–3424 (2005).
39. Craggs, J. K., Ball, J. K., Thomson, B. J., Irving, W. L. & Grabowska, A. M. Development of a strand-specific RT-PCR based assay to detect the replicative form of hepatitis C virus RNA. *J. Virol. Methods* **94,** 111–120 (2001).
40. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59,** 307–321 (2010).
41. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17,** 754–755 (2001).
42. Rambaut, A. & Drummond, A. J. Tracer v1.5 http://beast.bio.ed.ac.uk/Tracer (30 November 2009).
43. de Bruyn, M. *et al.* Paleo-drainage basin connectivity predicts evolutionary relationships across three southeast Asian biodiversity hotspots. *Syst. Biol.* **62,** 398–410 (2013).
44. Therneau, T. Coxme: Mixed Effects Cox Models http://CRAN.R-project.org/package=coxme (15 May 2012).
45. Reiczigel, J., Foldi, J. & Ozsvari, L. Exact confidence limits for prevalence of a disease with an imperfect diagnostic test. *Epidemiol. Infect.* **138,** 1674–1678 (2010).
46. Blaker, H. Confidence curves and improved exact confidence intervals for discrete distributions. *Can. J. Statist.* **28,** 783–798 (2000).
47. epiR:. an R package for the analysis of epidemiological data v. R package version 0.9-45 (30 November 2012).
48. Rossi, R. E., Mulla, D. J., Journel, A. G. & Franz, E. H. Geostatistical tools for modeling and interpreting ecological spatial dependence. *Ecol. Monogr.* **62,** 277–314 (1992).
49. Larmarange, J., Vallo, R., Yaro, S., Msellati, P. & Meda, N. Methods for mapping regional trends of HIV prevalence from demographic and health surveys (DHS). *Cybergeo.* http://cybergeo.revues.org/24606 (2011).
50. Moran, P. A. Notes on continuous stochastic phenomena. *Biometrika* **37,** 17–23 (1950).
51. Gittleman, J. L. & Kot, M. Adaptation: statistics and a null model for estimating phylogenetic effects. *Syst. Biol.* **39,** 227–241 (1990).
52. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20,** 289–290 (2004).
53. Baayen, R. H., Davidson, D. J. & Bates, D. M. Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* **59,** 390–412 (2008).
54. Bates, D., Maechler, M. & Bolker, B. lme4: Linear mixed-effects models using S4 classes (22 June 2012).
55. Schielzeth, H. Simple means to improve the interpretability of regression coefficients. *Meth. Ecol. Evol.* **1,** 103–113 (2010).
56. R Foundation for Statistical Computing R: a language and environment for statistical computing (26 October 2012).

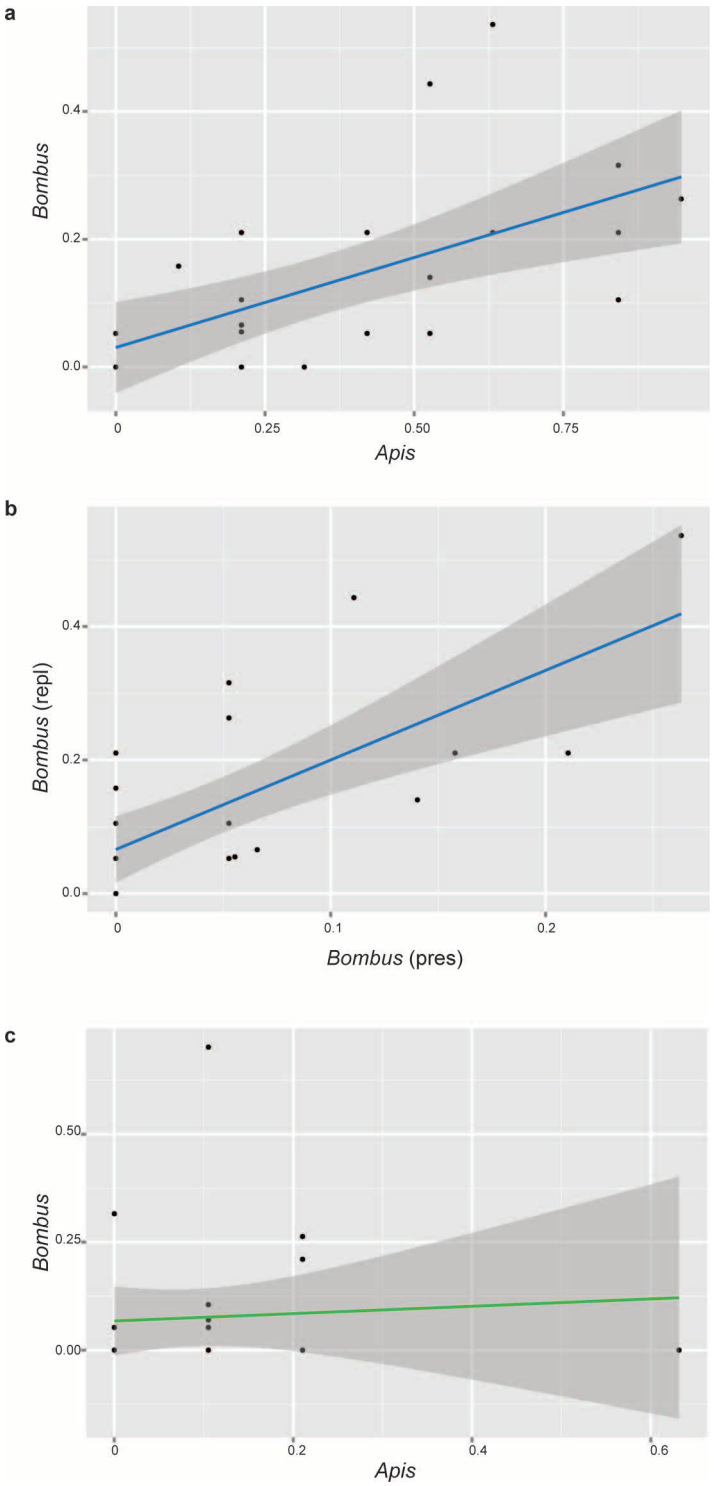


	Apis	*Bombus1*	*Bombus2*
A	10	10	10
B	10	10	10
C	10	10	10
D	10	10	10
E	10	9	10
F	10	10	10
G	0	6	10
H	10	10	10
I	10	10	10
J	10	9	10
K	10	10	10
L	10	2	10
M	10	10	10
N	10	10	10
O	10	9	10
P	10	6	10
Q	10	9	10
R	10	10	5
S	10	10	10
T	10	10	10
U	10	10	10
V	10	10	10
W	10	10	10
X	10	10	10
Y	10	10	10
Z	10	10	10

Extended Data Figure 1 | Host bee species and sampling-site distributions. Distribution of sampling sites across Great Britain and the Isle of Man. The most common *Bombus* species on a given site is represented by coloured letters and the second most common *Bombus* species is represented by the colours of the dots. Total sample sizes for each site are given in the table.



Extended Data Figure 2 | Prevalence of DWV and *N. ceranae* per site and host bee species. **a**–**d**, Pathogen prevalence in *Bombus* spp. in per cent per site for DWV (**a**) and for *N. ceranae* (**b**), and per species for DWV (**c**) and for *N. ceranae* (**d**). Bars indicate 95% confidence intervals. Note different scales.



Extended Data Figure 3 | Raw data for prevalence of DVW and *N. ceranae*. The linear models shown only illustrate the relationships but do not drive the conclusions in the main text. **a**, DWV presence in *Apis* and *Bombus* (adjusted $R^2 = 0.34$, $P < 0.001$). **b**, DWV replicating in *Bombus* and DWV presence in *Bombus* (adjusted $R^2 = 0.46$, $P < 0.001$). **c**, *N. ceranae* presence in *Apis* and *Bombus* (adjusted $R^2 = -0.04$, $P > 0.728$). The line shows the best fit and the dark grey region shows 95% confidence interval of fit.

Extended Data Table 1 | Pathogen prevalence per species

species (*N*)	*Apis* (250)	*B.ter* (170)	*B.luc* (60)	*B.lap* (175)	*B.pas* (60)	*B.hor* (20)	*B.mon* (10)
DWV present	36 [30, 43]	9 [5, 14]	18 [9, 29]	16 [11, 23]	4 [1, 12]	0 [0, 17]	11 [1, 47]
DWV replicating	88 [70, 98]*	1 [0, 3]	4 [1, 12]	10 [6, 15]	0 [0, 6]	0 [0, 17]	11 [1, 47]
N. ceranae	9 [6, 13]	2 [1, 6]	0 [0, 6]	16 [11, 23]	0 [0, 6]	5 [0, 25]	0 [0, 29]
single infection	18 [14, 23]	3 [1, 7]	3 [1, 12]	20 [14, 27]	0 [0, 6]	5 [0, 25]	11 [1, 47]
co-infection	1 [0, 3]	0 [0, 2]	0 [0, 6]	3 [1, 7]	0 [0, 6]	0 [0, 17]	0 [0, 29]

* The number given is out of the 31 DWV-present *Apis* samples tested. Pathogen prevalence is given in per cent with 95% confidence intervals [% prevalence, 95% confidence interval range]. Sample numbers (*N*) are shown in brackets.

Extended Data Table 2 | Alternative models for the diversification of DWV and VDV viruses in UK pollinators

Model		Marginal likelihood (ln)	Difference	BF	2 ln (BF)	Preferred model
Site (S)	Null	-1512.71	-94.92	$>10^{41}$	189.84	S
Host (H)		-1607.63				

The preferred tree model (Site) as determined by Bayes factor (BF) comparison.

LETTER

doi:10.1038/nature12871

Unidirectional pulmonary airflow patterns in the savannah monitor lizard

Emma R. Schachner[1], Robert L. Cieri[1], James P. Butler[2,3] & C. G. Farmer[1]

The unidirectional airflow patterns in the lungs of birds have long been considered a unique and specialized trait associated with the oxygen demands of flying, their endothermic metabolism[1] and unusual pulmonary architecture[2,3]. However, the discovery of similar flow patterns in the lungs of crocodilians indicates that this character is probably ancestral for all archosaurs—the group that includes extant birds and crocodilians as well as their extinct relatives, such as pterosaurs and dinosaurs[4–6]. Unidirectional flow in birds results from aerodynamic valves, rather than from sphincters or other physical mechanisms[7,8], and similar aerodynamic valves seem to be present in crocodilians[4–6]. The anatomical and developmental similarities in the primary and secondary bronchi of birds and crocodilians suggest that these structures and airflow patterns may be homologous[4–6,9]. The origin of this pattern is at least as old as the split between crocodilians and birds, which occurred in the Triassic period[10]. Alternatively, this pattern of flow may be even older; this hypothesis can be tested by investigating patterns of airflow in members of the outgroup to birds and crocodilians, the Lepidosauromorpha (tuatara, lizards and snakes). Here we demonstrate region-specific unidirectional airflow in the lungs of the savannah monitor lizard (*Varanus exanthematicus*). The presence of unidirectional flow in the lungs of *V. exanthematicus* thus gives rise to two possible evolutionary scenarios: either unidirectional airflow evolved independently in archosaurs and monitor lizards, or these flow patterns are homologous in archosaurs and *V. exanthematicus*, having evolved only once in ancestral diapsids (the clade encompassing snakes, lizards, crocodilians and birds). If unidirectional airflow is plesiomorphic for Diapsida, this respiratory character can be reconstructed for extinct diapsids, and evolved in a small ectothermic tetrapod during the Palaeozoic era at least a hundred million years before the origin of birds.

The lungs of lepidosaurs have been assumed to be ventilated tidally, an idea based on their bronchial architecture[11]; however, direct measurements of flow are lacking. Furthermore, fluid dynamics are often non-intuitive, and phenomena such as Venturi effects can result in complicated patterns of flow. It is important to characterize patterns of flow in lizards to assess the evolutionary history of the vertebrate lung. Varanids (73 species) are a widely distributed group of anguimorph lizards[12] with the oldest unambiguous fossil appearance of *Varanus* from the Upper Eocene and Lower Oligocene epochs of Egypt[13]. The external morphology of the genus *Varanus* is superficially conservative; however, they vary in mass by almost five orders of magnitude and occupy a range of ecological niches (semi-aquatic to arboreal)[14,15]. Compared to other lepidosaurs, varanids have high aerobic capacities, with *Varanus caudolineatus* having one of the highest rates of oxygen consumption ever recorded in a non-avian reptile[16]. High gas exchange capacities of varanids arise in part from their ability to supplement costal breathing with gular pumping[17]. These lizards possess multichambered (that is, multibronchial) lungs (Fig. 1a–e), which have long been used as a phylogenetic character for Varanoidea[18], and thus varanid pulmonary anatomy has received considerable attention[18–22].

Varanid lungs are large, structurally asymmetrical and multichambered[11,20] (Supplementary Video 1 and Fig. 1a–c). The dorsal surface is firmly attached to the ribs along most of their lengths[21]. The primary bronchus runs the length of the lung, ballooning into a large sac-like bronchus just distal to the ostium of the caudal-most lateral bronchus (as in Fig. 1b). The general arrangement of the secondary bronchi arising from the intrapulmonary primary bronchus follows what appears to be a stereotyped branching pattern. Without corroborating developmental data to support this observation, suppositions of bronchial identity remain tentative. Nevertheless, the anatomy of the adult bronchial tree can be visualized from computed tomography (CT) data of individual specimens of *V. exanthematicus*. The first bronchus to arise from the primary bronchus (the cervical bronchus, Cb) has a cartilaginous tube-shaped ostium that immediately makes a hairpin turn, running cranially and expanding into a large tubular bronchus terminating at the apex of the lung (Cb in Fig. 1c). This bronchus is anatomically reminiscent of the cervical ventral bronchus in crocodilians[6,9]. Arising sequentially from the lateral surface of the primary bronchus is a series of 9 to 11 variably shaped large sac-like bronchi, termed lateral bronchi (L1–10 in Fig. 1c).

Lateral bronchi communicate with the adjacent bronchi through numerous small intercameral perforations. A similar series of sequentially arranged sac-like bronchi arise off of the medial surface of the primary bronchus from small ostia and run caudomedially, rotating to a ventromedial position. Depending upon the individual, either the first or second medial bronchus on (usually) the right lung extends cranially along the ventral surface of the lung, terminating just distal to the carina. This bronchus is much smaller in the left lung and does not extend cranially. Along the dorsal surface of the primary bronchus in both lungs, small sac-like bronchi emerge in an asymmetrical pattern (Ssb in Fig. 1b). The respiratory parenchyma is largely restricted to the central and craniodorsal regions of the lung, with the saccular regions positioned at the cranial tip and caudoventral areas[21]. Small tertiary bronchi extend towards the pleural surface, forming hexagonal faveolar parenchyma[11,21].

The airflow patterns in *V. exanthematicus* are heterogeneous, with tidal and unidirectional flow observed in different regions of the lungs. Unidirectional airflow was measured using heated thermistor flow meters *in vivo* ($n = 5$) and in excised lungs ($n = 9$) in the caudal-most, large lateral bronchus (generally L10, depending upon the individual) (Fig. 1c, d and Fig. 2a–c). Flow was also observed by visualizing the movements of microspheres and pollen suspended in water in excised lungs ($n = 5$; Supplementary Video 2). Tidal flow was recorded in a cranial lateral bronchus (excised, $n = 2$), whereas unidirectional flow was observed visually in the rest of the lateral bronchi (L2–L10; $n = 4$) and the cervical bronchus ($n = 4$). Biased flow (that is, a significantly stronger magnitude of flow present during only one phase of the respiratory cycle) was measured in the abdominal sac-like bronchus (with thermistors ($n = 4$) and microspheres ($n = 4$) in excised lungs), with air arriving dorsally via the primary bronchus during inspiration, travelling caudoventrally to the caudal surfaces of the last lateral bronchus and through

[1]Department of Biology, University of Utah, Salt Lake City, Utah 84112, USA. [2]Division of Sleep Medicine, Department of Medicine, Harvard Medical School, Boston, Massachusetts 02215, USA. [3]Molecular and Integrative Physiologic Science Program, Department of Environmental Health, Harvard School of Public Health, Boston, Massachusetts 02115, USA.

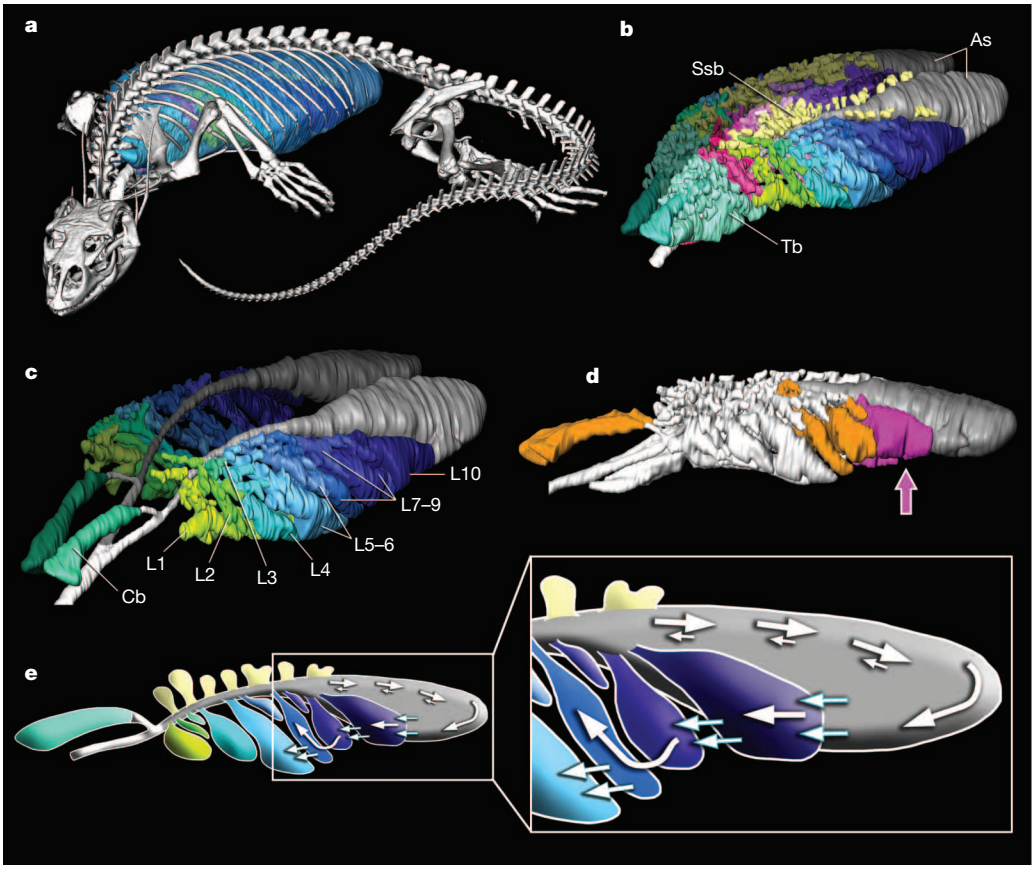


Figure 1 | Pulmonary anatomy and airflow patterns of *Varanus exanthematicus*. **a**, Volume rendered skeleton and segmented lungs. **b**, Solid representation of the bronchial tree. Tb, tertiary bronchi; Ssb, secondary sac-like bronchi; As, abdominal sac-like bronchus. **c**, Same as **b** with tertiary and medial bronchi removed. Cb, cervical bronchus; L1–L10, lateral bronchi 1–10. **d**, Bronchi in which flow was measured in excised lungs (orange/grey), and *in vivo* in this animal (pink). **e**, A diagram of the lung with arrows showing the direction of measured airflow (L5, L7, L9, L10 and abdominal sac). Paired large and small arrows indicate biased flow; blue arrows indicate interbronchial flow.

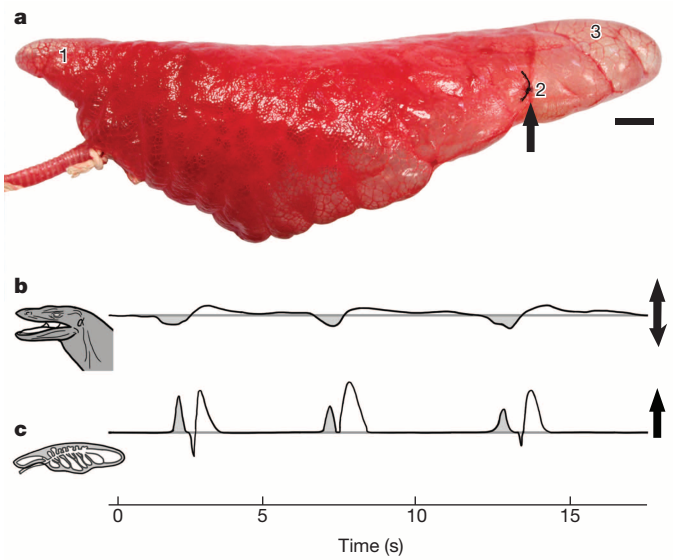


Figure 2 | Airflow recorded *in vivo*. **a**, Left lateral view of the left excised and inflated lung of *Varanus exanthematicus*. The arrow indicates where the airflow probe was surgically implanted for all *in vivo* measurements. Numbers represent regions where flow was recorded in excised lungs: 1, cervical bronchus (excised: $n = 2$; saline: $n = 4$); 2, last lateral bronchus (*in vivo*: $n = 5$; excised: $n = 9$; saline: $n = 4$); 3, abdominal sac (excised: $n = 4$, saline: $n = 4$). Scale bar = 1 cm. **b**, Tidal airflow measured at the nares. Shaded regions show inhalation; unshaded regions show exhalation. **c**, Largely unidirectional flow in L10; flows indicate directionality, not actual magnitudes.

the intercameral perforations along the shared bronchial walls during expiration (Fig. 1e). Flow between the lateral bronchi is interbronchial unidirectional flow, meaning that the flow travels in the same direction from one bronchus to another during both phases of the respiratory cycle.

These observations thus constitute evidence supporting the conclusion that the global pattern of flow in the bronchial tree of *V. exanthematicus* is predominantly unidirectional, composed of cranial and caudal regions of unidirectional airflow (Fig. 1e). Flow in the cervical bronchus appears to maintain its one-way direction by jetting in association with the anatomy of the ostium and proximally constricted bronchus, coupled with connections to small tertiary bronchi (Tb in Fig. 1b). Flow between the lateral bronchi is probably also maintained via jetting, in conjunction with the branching angle of each individual bronchus relative to the primary bronchus and the proximal constriction of each bronchus, thus constituting aerodynamic valving. The flow between bronchi is possible because of intercameral perforations, much like those found in archosaur lungs, despite the differences in their respective bronchial architecture. Aerodynamic valves arise from the geometry and branching angles of the primary and secondary bronchi, with valving mediated by the convective momentum of gas flow in this particular geometry. We suggest that this is the mechanism biasing flow in one direction in the lung of *V. exanthematicus*, owing to the absence of any physical flaps or muscular sphincters within their bronchial tree, and because flow patterns were unchanged in excised lungs. Thus it appears that the flow arises from the same aerodynamic phenomena seen in the archosaurian lung[7].

Unidirectional flow patterns have been measured in both avian and crocodilian lungs, indicating that this trait is probably plesiomorphic for Archosauria. The presence of unidirectional flow in regions of the

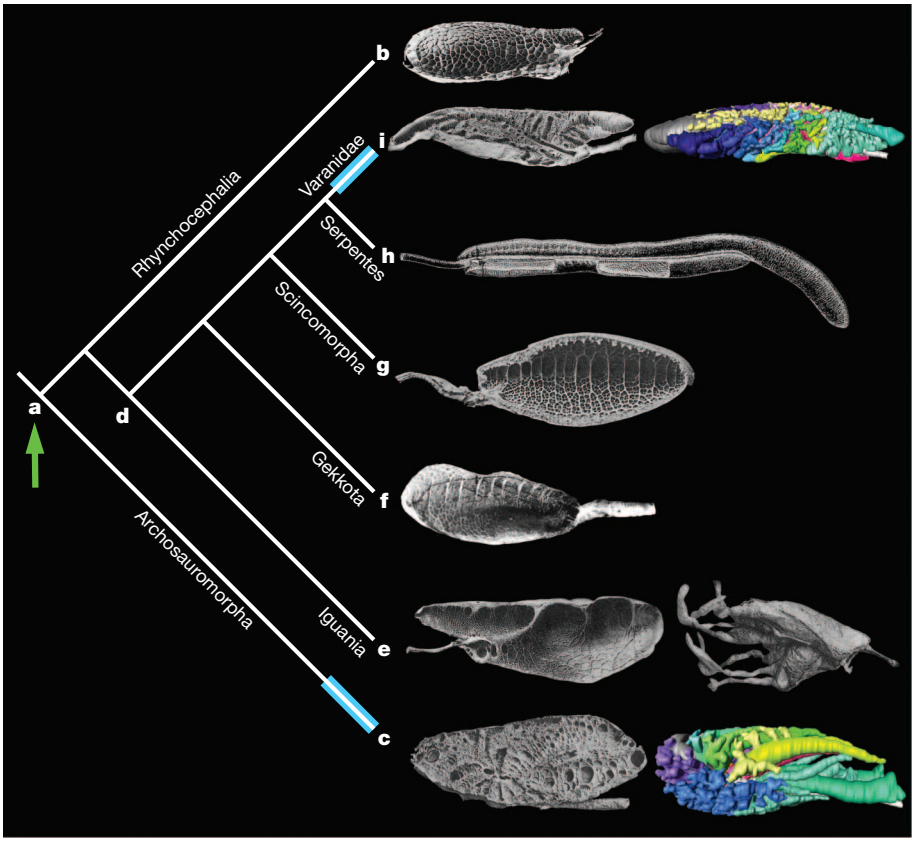


Figure 3 | Phylogeny for Diapsida showing lungs of representative taxa. Greyscale images are modified from Milani[24,25] and transected. The coloured three-dimensional images are the bronchial tree (right lateral view). Images are not to scale. **a**, Diapsida. **b**, *Sphenodon punctatus*. **c**, Crocodile sp. (left) and *Alligator mississippiensis* (right). **d**, Squamata. **e**, *Iguana iguana* (left) and *Polychrus marmoratus* (right). **f**, *Gekko gecko*[26]. **g**, *Lacerta viridis*. **h**, *Python sp.* in dorsal view[23] . **i**, *Varanus bengalensis* (left) and *V. exanthematicus* (right). The blue regions of the phylogeny reflect the hypothesis that unidirectional airflow evolved convergently; the green arrow shows the alternative hypothesis of an ancestral origin.

lung of *V. exanthematicus* thus raises two hypotheses reflecting different evolutionary scenarios (Fig. 3). The first hypothesis is that unidirectional flow patterns evolved independently in both Archosauria and Lepidosauria, and are a convergent apomorphy of both groups. The alternative hypothesis is that this pulmonary character is homologous in archosaurs and *V. exanthematicus*, having evolved only once, and is thus the ancestral state for diapsids (Lepidosauria + Archosauria). Relative to other lepidosaurs, varanids are particularly derived with a substantially more complex bronchial arrangement than their more basal relatives (see Fig. 3)[11,23]. The structure of varanid airways is different from that of archosaurs, both in terms of bronchial geometry as well as its branching sequence along the primary bronchus, making it difficult to favour one evolutionary scenario over the other. The presence of archosaur-like aerodynamic valves and unidirectional flow in the varanid lung suggests that this trait evolved in ancestral diapsids. However, differences in the patterns of flow between varanids (caudal to cranial in the ventrolateral sac-like bronchi) and archosaurs (caudal to cranial in the dorsal tube-shaped secondary bronchi), coupled with notable differences in the arrangement of the secondary bronchi, can be interpreted as convergent until the airflow patterns in more basal lepidosaurs are measured (Fig. 3). Owing to the considerable amount of variability in the lepidosaurian respiratory system, both across the entire clade as well as within individual groups, it will be essential to investigate flow patterns in multiple representative species from each major group (for example, *Sphenodon*, Iguania, Gekkota, Scincomorpha and other non-varanid anguimorphs) to shed light on this question (Fig. 3).

Determining when unidirectional airflow patterns first evolved has implications for understanding both the origin and function of respiratory patterns in non-avian reptiles, as well as reconstructing lung physiology and morphology in extinct taxa. If demonstrated to be ancestral for Diapsida, unidirectional airflow patterns can be parsimoniously reconstructed in all extinct diapsids. If these flow patterns evolved convergently between varanids and archosaurs, then this would suggest that the ability to ventilate unidirectionally holds an adaptive significance to these taxa.

METHODS SUMMARY

In vivo data were collected from five live specimens of *Varanus exanthematicus* (mass 350 g–930 g), and *ex vivo* data from ten specimens. Animals were acquired from the California Zoological Supply (live) and donated by J. Dix, Utah's Reptile Rescue Service (deceased). All experiments were performed in accordance with and approved by the University of Utah Institutional Animal Care and Use Committee. Three individuals were CT-scanned at 100 peak kilovolts and 400 milliamp tube current. A series of images were made along the long axis of the lungs. The thickness of each image (slice) was 0.6 mm and the slices were made at intervals of 0.4 mm along the long axis such that 0.2 mm of each slice overlapped with the previous slice. Digital models of the bronchial tree were segmented using Avizo version 7.1 (http://www.vsg3d.com/avizo/standard). Measurements of airflow were made with dual heated thermistor airflow probes surgically implanted into individual bronchi. The probes were connected to an HEC 132C Thermistor Flowmeter (Hector Engineering). The analogue output was converted to a digital signal (Biopac Systems) and recorded on a computer using AcqKnowledge software (Biopac Systems). Airflow at the nares was measured with a pneumotach (Hans Rudolph Inc.). Flow traces in live animals were recorded during hypercapnic breathing; traces measured in excised lungs were acquired from artificial ventilation (60 cm^3 syringe). Videos of the movement of saline containing microspheres (222 μm in diameter, Thermo Scientific) and pollen through excised lungs were taken with a Canon EOS T2i (resolution of 1,080 pixels) digital camera. The raw CT data are available from the Dryad Digital Repository at http://doi.org/10.5061/dryad.v1d30.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 August; accepted 6 November 2013.
Published online 11 December 2013.

1. Maina, J. N. Development, structure, and function of a novel respiratory organ, the lung-air sac system of birds: to go where no other vertebrate has gone. *Biol. Rev. Cambr. Phil. Soc.* **81,** 545–579 (2006).
2. Brackenbury, J. H. Lung-air-sac anatomy and respiratory pressures in the bird. *J. Exp. Biol.* **57,** 543–550 (1972).
3. Maina, J. N. Spectacularly robust! Tensegrity principle explains the mechanical strength of the avian lung. *Respir. Physiol. Neurobiol.* **155,** 1–10 (2007).
4. Farmer, C. G. The provenance of alveolar and parabronchial lungs: insights from paleoecology and the discovery of cardiogenic, unidirectional airflow in the American alligator (*Alligator mississippiensis*). *Physiol. Biochem. Zool.* **83,** 561–575 (2010).
5. Farmer, C. G. & Sanders, K. Unidirectional airflow in the lungs of alligators. *Science* **327,** 338–340 (2010).
6. Schachner, E. R., Hutchinson, J. R. & Farmer, C. G. Pulmonary anatomy in the Nile crocodile and the evolution of unidirectional airflow in Archosauria. *PeerJ* http://dx.doi.org/10.7717/peerj.60 (2013).
7. Butler, J. P., Banzett, R. B. & Fredberg, J. J. Inspiratory valving in avian bronchi: aerodynamic considerations. *Respir. Physiol.* **72,** 241–255 (1988).
8. Hazelhoff, E. H. Structure and function of the lung of birds. *Poult. Sci.* **30,** 3–10 (1951).
9. Sanders, R. K. & Farmer, C. G. The pulmonary anatomy of *Alligator mississippiensis* and its similarity to the avian respiratory system. *Anat. Rec.* **295,** 699–714 (2012).
10. Nesbitt, S. J. The early evolution of archosaurs: relationships and the origin of major clades. *Bull. Am. Mus. Nat. Hist.* **352,** 1–292 (2011).
11. Perry, S. F. in *Biology of the Reptilia* Vol. 19 (Morphology G) (eds Gans, C. & Gaunt, A. S.) 1–92 (Society for the Study of Amphibians and Reptiles, 1998).
12. Conrad, J. L., Balcarcel, A. M. & Mehling, C. M. Earliest example of a giant monitor lizard (*Varanus,* Varanidae, Squamata). *PLoS ONE* **7,** e41767 (2012).
13. Holmes, R. B., Murray, A. M., Attia, Y. S., Simons, E. L. & Chatrath, P. Oldest known Varanus (Squamata: Varanidae) from the Upper Eocene and Lower Oligocene of Egypt: support for an African origin of the genus. *Palaeontology* **53,** 1099–1110 (2010).
14. Collar, D. C., Schulte, J. A. II & Losos, J. B. Evolution of extreme body size disparity in monitor lizards (*Varanus*). *Evolution* **65,** 2664–2680 (2011).
15. Pianka, E. R. Evolution of body size: varanid lizards as a model system. *Am. Nat.* **146,** 398–414 (1995).
16. Thompson, G. G. & Withers, P. C. Standard and maximal metabolic rates of goannas (Squamata: Varanidae). *Physiol. Zool.* **70,** 307–323 (1997).
17. Owerkowicz, T., Farmer, C. G., Hicks, J. W. & Brainerd, E. L. Contribution of the gular pump to ventilation. *Science* **284,** 1661–1663 (1999).
18. Becker, H.-O., Böhme, W. & Perry, S. F. Die Lungenmorphologie der Warane (Reptilia: Varanidae) und ihre systematisch-stammesgeschichtliche Bedeutung. *Bonn. Zool. Beitr.* **40,** 27–56 (1989).
19. Burnell, A., Collins, S. & Young, B. A. The postpulmonary septum of *Varanus salvator* and its implication for Mosasaurian ventilation and physiology. *Bull. Soc. Geol. Fr.* **183,** 159–169 (2012).
20. Kirschfeld, U. Eine Bauplananalyse der Waranlunge. *Zool. Beitr. Neue Folge* **16,** 401–440 (1970).
21. Maina, J. N., Maloiy, G. M. O., Warui, C. N., Njogu, E. K. & Kokwaro, E. D. Scanning electron microscope study of the morphology of the reptilian lung: the savanna monitor lizard *Varanus exanthematicus* and the Pancake Tortoise *Malacochersus tornieri. Anat. Rec.* **224,** 514–522 (1989).
22. Perry, S. F. & Duncker, H. R. Lung architecture, volume and static mechanics in five species of lizards. *Respir. Physiol.* **34,** 61–81 (1978).
23. Wallach, V. in *Biology of the Reptilia* Vol. 19 (Morphology G) (eds Gans, C. & Gaunt, A. S.) 93–295 (Society for the Study of Amphibians and Reptiles, 1998).
24. Milani, A. Beiträge zur Kenntniss der Reptilienlunge. *Zool. Jahrb.* **7,** 545–592 (1894).
25. Milani, A. Beiträge zur Kenntnis der Reptilienlunge. II. *Zool. Jahrb.* **10,** 93–156 (1897).
26. Milsom, W. K. & Vitalis, T. Z. Pulmonary mechanics and the work of breathing in the lizard, *Gekko gecko. J. Exp. Biol.* **113,** 187–202 (1984).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Dix (Reptile Rescue Service) for the donation of deceased varanid specimens, J. Bourke for assistance with Avizo, and D. Shafer for German translations. This work was supported by an American Association of Anatomists Postdoctoral Fellowship and an American Philosophical Society Franklin Research Grant to E.R.S., National Science Foundation grants to C.G.F. (IOS-1055080 and IOS-0818973) and a generous donation to the Farmer laboratory by S. Meyer.

Author Contributions E.R.S. and R.L.C. conducted the *in vivo* surgeries. All authors collected data on excised lungs. E.R.S. acquired the CT scans and generated the three-dimensional digital models. C.G.F. and J.P.B. supervised and contributed ideas throughout the project. All authors contributed to the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.R.S. (eschachner@gmail.com) or C.G.F. (cg.frmr@gmail.com).

METHODS

Twelve animals were used in this study, and were acquired from the California Zoological Supply (live) and J. Dix, Utah's Reptile Rescue Service (deceased). No animals were excluded from the analysis. No randomization or blinding was done, and no statistical tests were used in this study. The animals were all *Varanus exanthematicus,* of largely unknown gender and age (mass 350 g–930 g). *In vivo* data were collected from five live specimens of *V. exanthematicus*. Data were collected from the excised lungs of ten specimens of mixed sex and unknown age. All experiments were performed in accordance with and approved by the University of Utah Institutional Animal Care and Use Committee. Three individuals were CT-scanned at 100 peak kilovoltage and 400 millamp tube current. A series of images were made along the long axis of the lungs. The thickness of each image (slice) was 0.6 mm and the slices were made at intervals of 0.4 mm along the long axis such that 0.2 mm of each slice overlapped with the previous slice. Digital models of the bronchial tree, lung surface and skeleton were segmented by hand in Avizo version 7.1 (http://www.vsg3d.com/avizo/standard) using a Wacom Intuos4 pen tablet. The images were edited into figures in Adobe Photoshop CS6, and the three-dimensional files exported from Avizo were edited into a video file in Adobe Premiere CS6. Measurements of airflow were made with dual heated thermistor airflow probes surgically implanted into individual bronchi of the lungs. The probes were connected to an HEC 132C Thermistor Flowmeter (Hector Engineering). The analogue output was converted to a digital signal (Biopac Systems) and recorded on a laptop using AcqKnowledge software (Biopac Systems). Airflow at the nares was measured with a pneumotach (Hans Rudolph Inc.). Flow traces in live animals were recorded as they breathed hypercapnic gas; traces measured in excised lungs were acquired from artificial ventilation (60 cm^3 syringe). Five of the smaller lungs were excised and filled with microsphere (222 μm in diameter, Thermo Scientific)-infused saline and pollen grains. Video of movement of the microspheres was recorded with a Canon EOS T2i (resolution of 1,080 pixels) digital camera. The raw CT data are available from the Dryad Digital Repository at http://doi.org/10.5061/dryad.v1d30.

LETTER

doi:10.1038/nature12881

Landscape of genomic alterations in cervical carcinomas

Akinyemi I. Ojesina[1,2]*, Lee Lichtenstein[2]*, Samuel S. Freeman[2], Chandra Sekhar Pedamallu[1,2], Ivan Imaz-Rosshandler[3], Trevor J. Pugh[1,2], Andrew D. Cherniack[2], Lauren Ambrogio[2], Kristian Cibulskis[2], Bjørn Bertelsen[4], Sandra Romero-Cordoba[3], Victor Treviño[5], Karla Vazquez-Santillan[3], Alberto Salido Guadarrama[3], Alexi A. Wright[1,6], Mara W. Rosenberg[2], Fujiko Duke[1], Bethany Kaplan[1,2], Rui Wang[1,7], Elizabeth Nickerson[2], Heather M. Walline[8], Michael S. Lawrence[2], Chip Stewart[2], Scott L. Carter[2], Aaron McKenna[2], Iram P. Rodriguez-Sanchez[9], Magali Espinosa-Castilla[3], Kathrine Woie[10], Line Bjorge[10,11], Elisabeth Wik[10,11], Mari K. Halle[10,11], Erling A. Hoivik[10,11], Camilla Krakstad[10,11], Nayeli Belem Gabiño[3], Gabriela Sofia Gómez-Macías[9], Lezmes D. Valdez-Chapa[9], María Lourdes Garza-Rodríguez[9], German Maytorena[12], Jorge Vazquez[12], Carlos Rodea[12], Adrian Cravioto[12], Maria L. Cortes[2], Heidi Greulich[1,2,6], Christopher P. Crum[13], Donna S. Neuberg[14], Alfredo Hidalgo-Miranda[3], Claudia Rangel Escareno[3,15], Lars A. Akslen[4,16], Thomas E. Carey[17], Olav K. Vintermyr[4,16], Stacey B. Gabriel[2], Hugo A. Barrera-Saldaña[9], Jorge Melendez-Zajgla[3], Gad Getz[2,18], Helga B. Salvesen[10,11]* & Matthew Meyerson[1,2,13]*

Cervical cancer is responsible for 10–15% of cancer-related deaths in women worldwide[1,2]. The aetiological role of infection with high-risk human papilloma viruses (HPVs) in cervical carcinomas is well established[3]. Previous studies have also implicated somatic mutations in *PIK3CA, PTEN, TP53, STK11* and *KRAS*[4–7] as well as several copy-number alterations in the pathogenesis of cervical carcinomas[8,9]. Here we report whole-exome sequencing analysis of 115 cervical carcinoma–normal paired samples, transcriptome sequencing of 79 cases and whole-genome sequencing of 14 tumour–normal pairs. Previously unknown somatic mutations in 79 primary squamous cell carcinomas include recurrent E322K substitutions in the *MAPK1* gene (8%), inactivating mutations in the *HLA-B* gene (9%), and mutations in *EP300* (16%), *FBXW7* (15%), *NFE2L2* (4%), *TP53* (5%) and *ERBB2* (6%). We also observe somatic *ELF3* (13%) and *CBFB* (8%) mutations in 24 adenocarcinomas. Squamous cell carcinomas have higher frequencies of somatic nucleotide substitutions occurring at cytosines preceded by thymines (Tp*C sites) than adenocarcinomas. Gene expression levels at HPV integration sites were statistically significantly higher in tumours with HPV integration compared with expression of the same genes in tumours without viral integration at the same site. These data demonstrate several recurrent genomic alterations in cervical carcinomas that suggest new strategies to combat this disease.

The prevention of cervical cancer by Pap smear-based screening and treatment programs has been largely successful in resource-rich countries. However, cervical cancer is the second most common cause of cancer-related deaths in women in developing countries, in which many patients are diagnosed at advanced stages of disease with limited treatment options and poor prognosis[1]. Recent advances in targeted therapy against specific somatic alterations have transformed the management of cancers in general[10], and the discovery of new therapeutic targets in cervical cancer could improve upon current strategies to combat cervical carcinomas.

To provide comprehensive data on the landscape of genomic aberrations that contribute to cervical cancer, we investigated a cohort that included 100 patients from Norway and 15 patients from Mexico (Supplementary Notes 1–7). We performed exome sequencing of 193,094 exons, covering a median of 34.2 megabases (Mb) at a median of 89× (range: 56–122×) coverage for tumour samples and 88× (range: 69–122×) coverage for normal samples, followed by calling of somatic mutations using the MuTect algorithm[11], and identified a total of 17,795 somatic mutations across the entire data set, including 11,419 missense, 936 nonsense, 4,643 silent, 219 splice site and 29 translation start site mutations, as well as 401 deletions and 131 insertions.

The aggregate nonsilent mutation rate across the data set was 3.7 per Mb. However, squamous cell carcinomas had a higher rate of nonsilent mutations (4.2 mutations per Mb) than adenocarcinomas (1.6 mutations per Mb) (Wilcoxon $P = 0.0095$). The clinical, pathological, epidemiological and mutational characteristics of the tumours are summarized in Supplementary Figs 1–6, Supplementary Tables 1–5 and Supplementary Notes 8 and 9.

Hierarchical clustering of all 115 tumours on the basis of mutational context revealed that most tumours were characterized by previously described[12] mutational signatures: with predominantly Tp*C-to-T/G mutations and *CpG-to-T mutations (Fig. 1 and Supplementary Fig. 4). Tp*C mutations were present at a relative frequency of >0.5 in 53 (46%) tumours, and the relative frequency of Tp*C mutations was positively correlated with mutation rates, especially in squamous cell carcinomas (Fig. 1, Supplementary Note 8 and Supplementary Fig. 5). In addition, 5,648 (54%) of the 10,328 nonsilent mutations observed in squamous cell carcinomas were Tp*C-to-T/G mutations.

We performed mutation significance analyses on 79 squamous cell carcinomas and 24 adenocarcinomas. Genes were determined to be significantly mutated if recurrent mutations were found in that gene at a false discovery rate of $q < 0.1$ after correction for multiple hypothesis testing, as described previously[13] (Supplementary Note 6). Details of candidate mutation validation are presented in Supplementary Figs 6 and 7.

As expected, recurrent mutations in *PIK3CA*, *PTEN* and *STK11* were present in 14%, 6% and 4%, respectively, of 79 squamous cell carcinomas (Table 1). In addition, we found significantly recurrent mutations in *EP300* (16%), *FBXW7* (15%), *HLA-B* (9%), *MAPK1* (8%) and *NFE2L2*

[1]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. [2]The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02142, USA. [3]Instituto Nacional de Medicina Genomica, Mexico City 14610, Mexico. [4]Department of Pathology, Haukeland University Hospital, N5021 Bergen, Norway. [5]Tecnológico de Monterrey, Monterrey 64849, Mexico. [6]Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. [7]Department of Thoracic Surgery, Fudan University Shanghai Cancer Center, Shanghai 200032, China. [8]Cancer Biology Program, Program in the Biomedical Sciences, Rackham Graduate School, University of Michigan, Ann Arbor, Michigan 48109, USA. [9]Facultad de Medicina y Hospital Universitario 'Dr. José Eluterio González' de la Universidad Autónoma de Nuevo León, Monterrey, Nuevo León 64460, México. [10]Department of Obstetrics and Gynecology, Haukeland University Hospital, N5021 Bergen, Norway. [11]Department of Clinical Science, Centre for Cancer Biomarkers, University of Bergen, N5020 Bergen, Norway. [12]Instituto Mexicano del Seguro Social, Mexico City 06720, Mexico. [13]Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. [14]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. [15]Claremont Graduate University, Claremont, California 91711, USA. [16]Centre for Cancer Biomarkers, Department of Clinical Medicine, University of Bergen, N5020 Bergen, Norway. [17]Head and Neck Oncology Program and Department of Otolaryngology, University of Michigan Comprehensive Cancer Center, Ann Arbor, Michigan 38109, USA. [18]Massachusetts General Hospital Cancer Center and Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA.
*These authors contributed equally to this work.

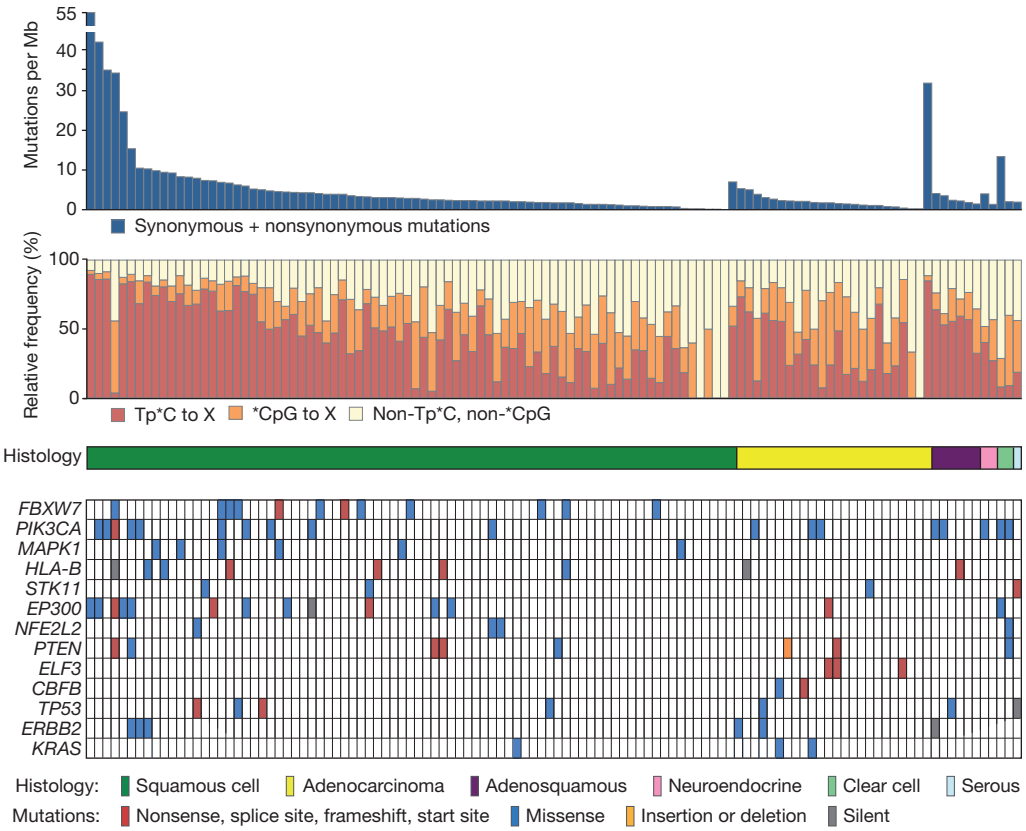


Figure 1 | Relationship of mutational spectrum and rates with clinicopathological characteristics in cervical carcinoma. All panels are aligned with vertical tracks representing 115 individuals. The data are sorted in order by histology (middle) and total mutational rate (top). The relative frequencies of nucleotide mutations occurring at cytosines preceded by thymines (Tp*C sites) or at cytosines followed by guanines (*CpG sites) are depicted in red and orange, respectively, in the middle panel. The bottom heatmap shows the distribution of mutations in significantly mutated genes ($q < 0.1$) in squamous cell carcinomas and adenocarcinomas in the order listed in Table 1. *TP53*, *ERBB2* and *KRAS* were significantly recurrent ($q < 0.1$) among cancer driver genes reported in COSMIC.

(4%), here reported for the first time, to our knowledge, in primary squamous cell cervical carcinomas (Table 1, Fig. 1, Supplementary Table 6 and Supplementary Fig. 8). In addition, *TP53* (9%) and *ERBB2* (5%) were found to be significantly mutated in analyses focused only on genes previously reported as mutated in the COSMIC database (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic) (Supplementary Table 8a). Interestingly, three out of the six *ERBB2* mutations (S310F, S310Y and V842I; Supplementary Fig. 8) are known oncogenic driver mutations and *in vitro* therapeutic targets in lung[14] and breast cancer[15].

Somatic *MAPK1* mutations were observed in 6 out of 79 squamous cell carcinomas of the cervix (7%), each involving a G-to-A transition resulting in recurrent E322K mutations in four individuals, and E81K and E220K mutations in one individual each (Fig. 2). To our knowledge, this is the first report of recurrent mutations of *MAPK1* in primary human cancer, although the *MAPK1* E322K mutation has been reported in an oropharyngeal carcinoma cell line[16] and scattered unpublished reports are summarized in COSMIC. The recurrent site-specific *MAPK1* mutations and the known role of the MAPK signalling pathway in cancer[17] suggest the possibility that mutant *MAPK1* may exert oncogenic activity.

We observed *EP300* and *FBXW7* mutations in our data set, similar to recent reports in endometrial and head and neck cancers[18,19]. Thirteen

Table 1 | Genes with significantly recurrent somatic mutations in cervical carcinomas

Gene	Description	Nonsilent mutations	Relative frequency (%)	Patients	Unique sites	Silent mutations	Indel + null	q
Squamous cell carcinoma ($n = 79$)								
*FBXW7**	F-box and WD repeat domain containing 7	12	15	12	8	0	2	4.03×10^{-12}
PIK3CA	Phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha	11	14	10	5	0	1	$<9.08 \times 10^{-12}$
*MAPK1**	Mitogen-activated protein kinase 1	6	8	6	3	0	0	0.000671
HLA-B‡	Major histocompatibility complex, class I, B	7	9	6	7	1	3	0.00169
STK11	Serine/threonine kinase 11	3	4	2	2	0	1	0.012
EP300‡	E1A binding protein p300	13	16	12	13	1	4	0.0354
NFE2L2‡	Nuclear factor, erythroid 2-like 2	3	4	3	2	0	0	0.0597
PTEN	Phosphatase and tensin homologue (mutated in multiple advanced cancers 1)	5	6	5	5	0	3	0.0693
Adenocarcinoma ($n = 24$)								
ELF3†	E74-like factor 3 (ets domain transcription factor, epithelial-specific)	3	13	3	3	0	3	0.03
CBFB†	core-binding factor, beta subunit	2	8	2	2	0	1	0.0342

Indel, insertions or deletions; null, nonsense, frameshift or splice-site mutations; *q*, *q* value, false discovery rate (Benjamini–Hochberg procedure).
* Genes with mutations observed in only squamous cell carcinomas.
† Genes with mutations observed in only adenocarcinomas.
‡ Genes with a majority of mutations occurring in squamous cell carcinomas.

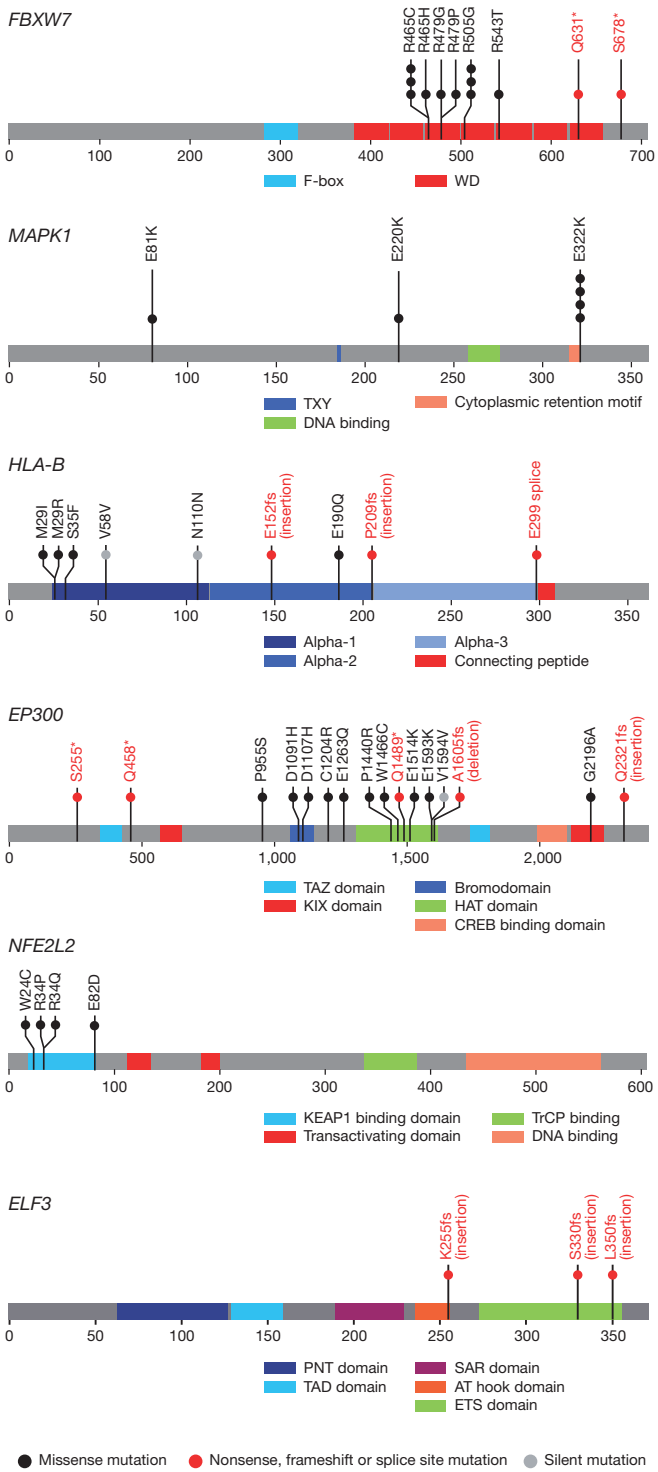


Figure 2 | Novel recurrent somatic mutations in cervical carcinoma. The locations of somatic mutations in novel significantly mutated genes in 115 cervical carcinoma. *FBXW7*, *MAPK1*, *HLA-B*, *EP300*, *NFE2L2* and *ELF3* are shown in the context of protein domain models derived from UniProt and Pfam annotations. Numbers refer to amino acid residues. Each filled circle represents an individual mutated tumour sample: missense and silent mutations are represented by filled black and grey circles, respectively, whereas nonsense, frameshift and splice site mutations are represented by filled red circles and red text. Domains are depicted with various colours with an appropriate key located below each domain model.

out of 15 nonsilent *EP300* mutations are novel in cancer, with eight of these (including two nonsense) residing in the histone acetyltransferase- and bromo-domains required for *EP300* activity[20] (Fig. 2). In addition, there were two truncating mutations at residues S255 and Q458 in *EP300*. The *FBXW7* gene also had two novel truncating mutations at residues Q631 and R678, with ten other mutations residing in the WD40 domains required to form the scaffold for the Skp1–Cul1–F-box protein complex[21]. Furthermore, all four *NFE2L2* mutations (W24C, R34P, R34Q and E82D) are in the domain required for interacting with its negative regulator, KEAP1 (ref. 22) (Fig. 2), consistent with similar findings in lung squamous cancers[23]. Interestingly, mutations in these genes (*FBXW7*, *EP300*, *MAPK1*, *NFE2L2*) occur largely in a non-overlapping pattern in our data set (Fig. 1 and Supplementary Fig. 6b). These observations suggest that epigenetic regulation and the oxidative stress response may have important roles in cervical cancer pathogenesis.

We found four missense and three frameshift mutations in the *HLA-B* gene encoding the histocompatibility leukocyte antigen HLA-B (Fig. 2 and Table 1). In addition, there were somatic mutations in other genes involved in antigen presentation, including splice site, nonsense and frameshift mutations in *HLA-A*, a gene previously reported as mutated in squamous cell carcinomas of the lung[23], and in the β_2-microglobulin (*B2M*) gene (Supplementary Fig. 8 and Supplementary Table 6). All mutations in these three genes were within the antigen-presenting domains of each respective protein[24]. Intriguingly, pathway analyses also revealed that the most significantly mutated gene set in squamous cell carcinomas involves immune response genes in the interferon-γ signalling pathway, including mutations in *IFNG* and *JAK2* (Supplementary Table 9a). Together, these data highlight the potential significance of the synergy between HPV infection and an altered immune response in the pathogenesis of squamous cell carcinomas of the cervix.

We also investigated a smaller subset of 24 adenocarcinomas. Our analysis revealed the *ELF3* (13%) and *CBFB* (8%) genes as recurrently mutated at $q < 0.1$ (Fig. 2, Table 1, Supplementary Fig. 8 and Supplementary Table 7). In addition, *PIK3CA* (16%) and *KRAS* (8%) were found to be significantly mutated in analyses focused only on genes previously reported as mutated in the COSMIC database (Supplementary Table 8b), consistent with previous reports[6]. Furthermore, gene set analyses revealed that the *PIK3CA/PTEN* pathway was significantly recurrently mutated across the adenocarcinoma subset (Supplementary Table 9b).

The *ELF3* mutations in the adenocarcinomas involve frameshift insertional events at amino acid positions 255, 330 and 350. *ELF3* (E74-like factor-3) encodes a member of the ETS transcription factor family that is expressed and upregulated in epithelial cancers, and is both a regulator and downstream effector of the ERBB2 signalling pathway[25]. Interestingly, *ELF3*-mutated tumours have higher gene expression levels than *ELF3* wild-type tumours (Supplementary Fig. 9).

The spectrum of somatic copy-number alterations, rearrangements, gene expression profiles, HPV integration and other genomic events observed in this cohort are documented in Supplementary Notes 10–14, Supplementary Tables 10–21 and Supplementary Figs 10–30. HPV integration sites were defined by the presence of least six chimaeric read pairs, derived from RNA sequencing (RNA-seq) data, in which the pairmate of an HPV sequence read mapped to the human genome (Supplementary Note 7). As expected, HPV integration sites were found within or in close proximity to several fragile sites (Supplementary Note 11) as well as previously reported genes[26–29] including *MYC*, *ERBB2*, *TP63*, *FANCC*, *RAD51B* and *CEACAM5* (Supplementary Table 14).

HPV integration occurred closer to amplified regions than expected by chance (Mann–Whitney $P < 2.2 \times 10^{-16}$; Fig. 3a), with 21 (41%) of 51 integration sites overlapping with amplified regions, supporting the hypothesis that viral integration may trigger genome amplification[30]. In general, viral integration was localized to one locus in most tumours investigated, and most of the integration sites were observed only in one tumour each (Supplementary Table 14). In addition, many of the genes involved in the integration events are members of cellular pathways known to have important roles in cancer (Supplementary Table 15). Similar to recent observations[29], we observed recurrent HPV integration into the *RAD51B* locus in three different tumours; intriguingly,

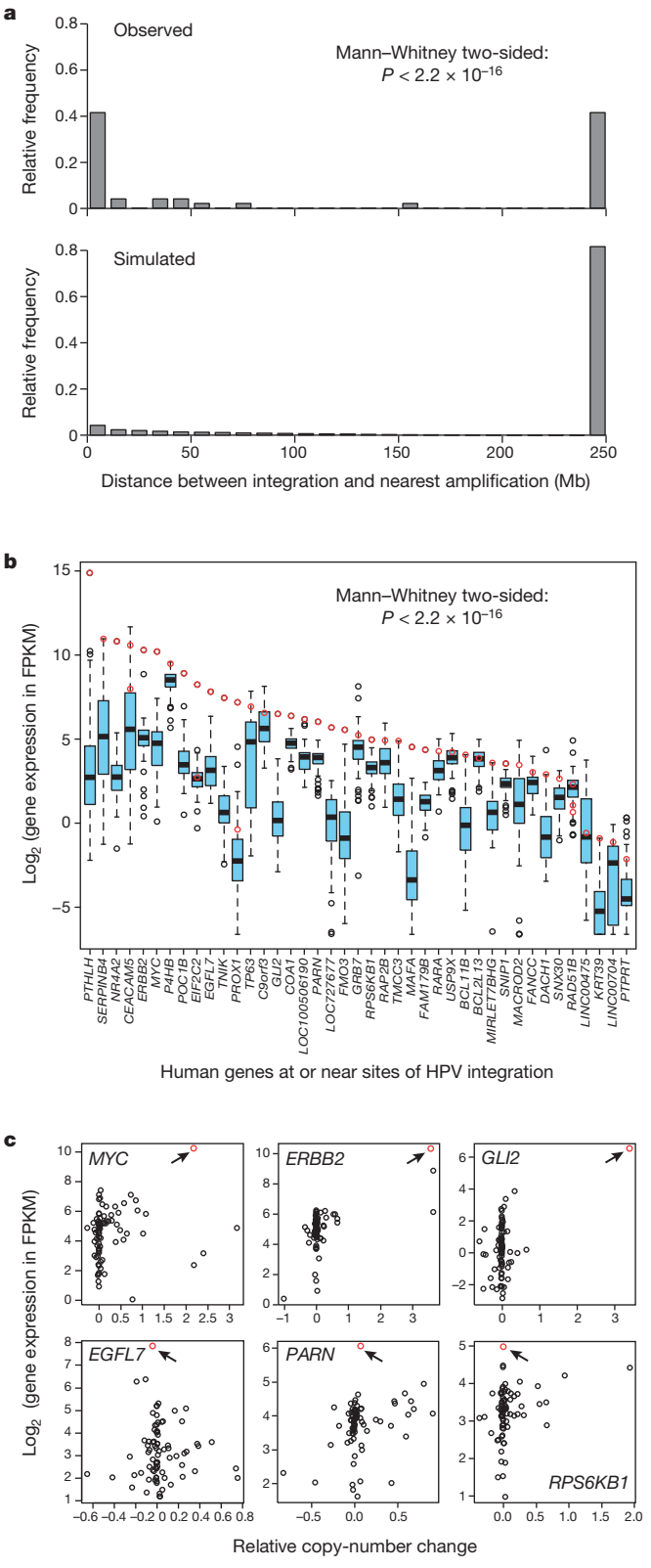


Figure 3 | Relationships between HPV integration, copy-number amplifications and gene expression in cervical carcinoma. **a**, Comparative histograms of true and simulated genomic distances (in Mb) between HPV integration sites and the nearest copy-number amplification (log segment mean difference >0.5). **b**, Boxplots of gene expression levels across 79 cervical tumours for 41 genes with chimaeric human–HPV read pairs. The expression levels for tumours with HPV integration in the respective genes are highlighted in red circles. FPKM, fragments per kilobase of exon per million fragments mapped. **c**, Scatter plots comparing copy-number alterations and gene expression levels across 79 tumours in selected integration site genes. The red circles represent data for the tumours with HPV integration events involving the respective genes.

each involved a different HPV type: HPV16, HPV18 and HPV52 (Supplementary Figs 19–21).

We also observed that gene expression levels at sites of HPV integration were significantly higher in tumours with HPV integration compared with the expression levels of the same genes across the other tumours without integration at that site ($P < 2.2 \times 10^{-16}$; Fig. 3b and Supplementary Figs 22–24). For some integration sites, including *MYC*, *ERBB2*, *GLI2*, *TNIK*, *NR4A2*, *PROX1*, *EIF2C2*, *FAM179B* and *SERPINB4*, high gene expression levels were associated with copy-number gains (Fig. 3c and Supplementary Fig. 25). Conversely, there were no copy-number changes at several other highly expressed integration sites including *RPS6KB1*, *MAFA*, *PARN*, *EGFL7*, *SNIP1*, *POC1B* and *BCL11B* (Fig. 3c, Supplementary Fig. 26 and Supplementary Notes 11E, F), supporting the hypothesis that the increased expression of these genes may be driven in part by integrated viral promoter(s)[27].

In summary, this study has demonstrated relationships between recurrent somatic mutations, copy-number alterations, gene expression and HPV integration in cervical carcinomas. We report significantly recurrent somatic mutations in the *MAPK1* gene in squamous cell cervical cancers, to our knowledge the first such report in human cancers. In addition, we found evidence of potential *ERBB2* activation by somatic mutation, amplification and HPV integration, suggesting that some cervical carcinoma patients could potentially be considered as candidates for clinical trials of ERBB2 inhibitors. Furthermore, our data suggest that alterations in immune response genes may synergize with HPV infection in the pathogenesis of squamous cell carcinomas. Finally, our data suggests that the association between HPV integration and increased expression of adjacent genes is a widespread phenomenon in primary cervical carcinomas.

METHODS SUMMARY

Following Institutional Review Board approvals and informed consent, nucleic acids were extracted from resected tumour tissue and normal peripheral blood in 115 patients. Exome, whole genome and cDNA sequencing was performed for 115, 14 and 79 patients, respectively, on Illumina HiSeq 2000 instruments. A flowchart of the entire analysis procedure can be found in Supplementary Fig. 1. A detailed description of the materials, analyses and validation is found in the Methods section.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 December 2012; accepted 13 November 2013.
Published online 25 December 2013.

1. Jemal, A. *et al.* Global cancer statistics. *CA Cancer J. Clin.* **61,** 69–90 (2011).
2. International Agency for Research on Cancer. A review of human carcinogen: biological agents. in *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans* Vol. 100B (International Agency for Research on Cancer, 2012).
3. zur Hausen, H. Papillomaviruses in the causation of human cancers — a brief historical account. *Virology* **384,** 260–265 (2009).
4. Crook, T. *et al.* Clonal p53 mutation in primary cervical cancer: association with human-papillomavirus-negative tumours. *Lancet* **339,** 1070–1073 (1992).
5. McIntyre, J. B. *et al. PIK3CA* mutational status and overall survival in patients with cervical cancer treated with radical chemoradiotherapy. *Gynecol. Oncol.* **128,** 409–414 (2013).
6. Kang, S. *et al.* Inverse correlation between *RASSF1A* hypermethylation, *KRAS* and *BRAF* mutations in cervical adenocarcinoma. *Gynecol. Oncol.* **105,** 662–666 (2007).
7. Wingo, S. N. *et al.* Somatic *LKB1* mutations promote cervical cancer progression. *PLoS ONE* **4,** e5137 (2009).
8. Narayan, G. & Murty, V. V. Integrative genomic approaches in cervical cancer: implications for molecular pathogenesis. *Future Oncol.* **6,** 1643–1652 (2010).
9. Vazquez-Mena, O. *et al.* Amplified genes may be overexpressed, unchanged, or downregulated in cervical cancer cell lines. *PLoS ONE* **7,** e32667 (2012).
10. Arteaga, C. L. & Baselga, J. Impact of genomics on personalized cancer medicine. *Clin. Cancer Res.* **18,** 612–618 (2012).
11. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnol.* **31,** 213–219 (2013).
12. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).

13. Lohr, J. G. *et al.* Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc. Natl Acad. Sci. USA* **109,** 3879–3884 (2012).
14. Greulich, H. *et al.* Functional analysis of receptor tyrosine kinase mutations in lung cancer identifies oncogenic extracellular domain mutations of *ERBB2*. *Proc. Natl Acad. Sci. USA* **109,** 14476–14481 (2012).
15. Bose, R. *et al.* Activating HER2 mutations in HER2 gene amplification negative breast cancer. *Cancer Discov.* **3,** 224–237 (2012).
16. Arvind, R. *et al.* A mutation in the common docking domain of ERK2 in a human cancer cell line, which was associated with its constitutive phosphorylation. *Int. J. Oncol.* **27,** 1499–1504 (2005).
17. De Luca, A., Maiello, M. R., D'Alessio, A., Pergameno, M. & Normanno, N. The RAS/RAF/MEK/ERK and the PI3K/AKT signalling pathways: role in cancer pathogenesis and implications for therapeutic approaches. *Expert Opin. Ther. Targets* **16** (Suppl. 2), S17–S27 (2012).
18. Le Gallo, M. *et al.* Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. *Nature Genet.* **44,** 1310–1315 (2012).
19. Agrawal, N. *et al.* Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in *NOTCH1*. *Science* **333,** 1154–1157 (2011).
20. Chen, J., Ghazawi, F. M. & Li, Q. Interplay of bromodomain and histone acetylation in the regulation of p300-dependent genes. *Epigenetics* **5,** 509–515 (2010).
21. Smith, T. F., Gaitatzes, C., Saxena, K. & Neer, E. J. The WD repeat: a common architecture for diverse functions. *Trends Biochem. Sci.* **24,** 181–185 (1999).
22. Tong, K. I. *et al.* Keap1 recruits Neh2 through binding to ETGE and DLG motifs: characterization of the two-site molecular recognition model. *Mol. Cell. Biol.* **26,** 2887–2900 (2006).
23. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519–525 (2012).
24. Pamer, E. & Cresswell, P. Mechanisms of MHC class I-restricted antigen processing. *Annu. Rev. Immunol.* **16,** 323–358 (1998).
25. Neve, R. M., Ylstra, B., Chang, C. H., Albertson, D. G. & Benz, C. C. ErbB2 activation of ESX gene expression. *Oncogene* **21,** 3934–3938 (2002).
26. Wentzensen, N., Vinokurova, S. & von Knebel Doeberitz, M. Systematic review of genomic integration sites of human papillomavirus genomes in epithelial dysplasia and invasive cancer of the female lower genital tract. *Cancer Res.* **64,** 3878–3884 (2004).
27. Kraus, I. *et al.* The majority of viral-cellular fusion transcripts in cervical carcinomas cotranscribe cellular sequences of known or predicted genes. *Cancer Res.* **68,** 2514–2522 (2008).
28. Schmitz, M., Driesch, C., Jansen, L., Runnebaum, I. B. & Durst, M. Non-random integration of the HPV genome in cervical cancer. *PLoS ONE* **7,** e39632 (2012).
29. Tang, K. W., Alaei-Mahabadi, B., Samuelsson, T., Lindh, M. & Larsson, E. The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nature Commun.* **4,** 2513 (2013).
30. Peter, M. *et al.* Frequent genomic structural alterations at HPV insertion sites in cervical carcinoma. *J. Pathol.* **221,** 320–330 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was conducted as part of the Slim Initiative for Genomic Medicine in the Americas, a project funded by the Carlos Slim Health Institute in Mexico. This work was also partially supported by the Rebecca Ridley Kry Fellowship of the Damon Runyon Cancer Research Foundation (A.I.O.); MMRF Research Fellow Award (A.I.O.); Helse Vest, Research Council of Norway, Norwegian Cancer Society and Harald Andersens legat (H.B.S.); CONACyT grant SALUD-2008-C01-87625 and UANL PAICyT grant CS1038-1 (H.A.B.-S.); and CONACyT grant 161619 (J.M.-Z.). We also thank B. Edvardsen, K. Dahl-Michelsen, Å. Mokleiv, K. Madisso, T. Njølstad and E. Valen for technical and programmatic assistance; the staff of the Broad Institute Genomics Platform for their assistance in processing samples and generating the sequencing data used in the analyses; the Instituto Mexicano del Seguro Social (IMSS) for their Support; and L. Gaffney of Broad Institute Communications for figure layout and design.

Author Contributions A.I.O., L.L., S.S.F., C.S.P., H.B.S. and M.M. wrote the manuscript with help from co-authors. A.I.O., L.L., K.C., C.S. and G.G. performed whole exome and genome sequencing data analysis. A.I.O., I.I., V.T., K.V.-S., A.S.G., S.R.-C., C.R.E., S.S.F. and C.S.P. performed RNA sequencing data analysis. A.I.O., S.S.F., C.S.P. and T.J.P. performed HPV integration analyses. A.I.O. and A.D.C. performed copy-number analyses. A.I.O., F.D., B.K., R.W. and H.G. performed functional experiments on *MAPK1*. B.B., N.B.G., G.S.G.-M. and C.P.C. facilitated and performed pathology review. O.K.V., H.M.W. and T.E.C. performed HPV status determination. L.A., E.N. and M.L.C. facilitated project management. L.L., I.I.-R., V.T., K.V.-S., A.S.G., S.R.-C., I.P.R.-S. and C.R.E. performed sequencing data validation. M.E.-C., M.K.H., E.W., E.A.H., C.K. and M.L.G.-R. performed specimen processing, biobanking and data management. K.W., L.B., L.D.V.-C., G.M., J.V., C.R., A.C. and H.B.S. collected patient materials and clinical information. A.I.O., L.L. and D.S.N. performed biostatistical and epidemiological analyses. A.I.O., L.L., S.S.F., C.S.P., I.I.-R., T.J.P., A.D.C., V.T., A.A.W., M.W.R., F.D., M.S.L., C.S., S.L.C., A.M., H.B.S. and M.M. contributed text, figures (including Supplementary Information) and analytical tools. A.H.-M., C.R.E., L.A.A., S.B.G., H.A.B.-S., J.M.-Z., G.G., H.B.S. and M.M. provided leadership for the project. All authors contributed to the final manuscript. Lead authors A.I.O. and L.L. and senior authors M.M. and H.B.S. contributed equally to this work.

Author Information Sequence data used for this analysis are available in dbGaP under accession phs000600. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.M. (matthew_meyerson@dfci.harvard.edu) or H.B.S. (helga.salvesen@uib.no).

METHODS

Sample preparation. All samples were obtained under Institutional Review Board approval and with documented informed consent. Surgically resected tumours or biopsies were snap frozen in liquid nitrogen and stored at −80 °C. Genomic DNA and RNA were extracted from tumours found by frozen section investigations to have >40% malignant epithelial cell component. A detailed description of sample collection is found in Supplementary Note 1. Nucleic acid was extracted using standard protocols described in Supplementary Note 2.

Sequence data generation. DNA from 115 tumour–normal paired samples was subjected to Agilent Sure-Select Human All Exon v2.0 based hybrid selection[31] followed by exome library construction for Illumina sequencing, and 14 pairs for whole genome library construction; complementary DNA from 79 samples was subjected to transcriptome library construction, according to standard methods. Twelve tumour–normal pairs were sequenced with all three types of libraries. All libraries were sequenced with the Illumina HiSeq 2000 instrument (Supplementary Note 4). Exome sequencing was performed using a hybrid capture of 193,094 exon targets from 18,862 coding genes. Reads were aligned to human genome build GRCh37 using a Burrows–Wheeler aligner[32]. Data from the Illumina HiSeq were converted into BAM files[33] (http://samtools.sourceforge.net/SAM1.pdf) for each sample, using Picard (http://picard.sourceforge.net/) (Supplementary Note 6A).

Variant calling and significance analysis. All variant calls and significance analysis were obtained using the standard Cancer Genome Analysis Pipeline[34–38] with some modifications detailed in Supplementary Note 6B. Cross-individual contamination was estimated using ContEst[39] with both single nucleotide polymorphism (SNP) Array (Supplementary Note 5) and sequencing data as input. Somatic single nucleotide variant (sSNV) and indel calls were generated using MuTect[11,34–36] and Indelocator[34–36], respectively, for all complete exome tumour–normal pairs. Variants were mapped to genes and transcripts using Oncotator (http://www.broadinstitute.org/oncotator), as well as being annotated with useful information such as overlapping COSMIC[40] records. D-ToxoG (http://www.broadinstitute.org/cancer/cga/dtoxog) was used to filter mutations generated by a sequencing artefact that was discovered during this project. Significantly mutated genes and gene sets (Supplementary Note 9) were identified by MutSig2.0 (ref. 13) on the basis of sSNV and indel calls. Mutation rate calculations were also provided by MutSig2.0. Rearrangements were identified using dRanger[37] running on data generated by whole genome sequencing (WGS) of 14 tumour–normal pairs (Supplementary Note 13A), on the basis of read pairings with unexpected distance or orientation. Somatic copy-number alterations, both broad and focal, were identified with the GISTIC 2.0 (refs 35, 36) tool, using segments identified from exome sequencing data (Supplementary Note 6B) as input (Supplementary Note 10A). Overlap between copy-number alterations and somatic nonsilent mutations, for a subset of significantly mutated genes, are reported in Supplementary Note 10D and Supplementary Fig. 14. UnifiedGenotyper[41,42] was used to identify germline mutations in genes in the Fanconi anaemia pathway and *TERC*, as reported in Supplementary Notes 14B, C. ABSOLUTE[43] was used to generate purity and ploidy estimates for tumour samples where SNP Array data was available (Source Data File 1).

Variant validation. Two validation approaches were used in this study. The first was to re-sequence mutations in significant genes. Libraries were constructed with 200 base pair flanks around key mutations in significant genes and sequenced using Illumina MiSeq. Manual validation was performed by examining mutations in the resulting BAM files. Mutations were considered validated if supported by five or more reads. The second validation approach was to compare exome sSNV calls against the corresponding WGS and/or RNA-seq calls where available. Mutations were considered validated if the alternate allele was seen in at least two reads and the calculated power was 80% or higher. These two approaches are detailed in Supplementary Note 6D. Owing to the variable nature of human leukocyte antigen (*HLA*) genes[44], mutations in *HLA-A* and *HLA-B* were validated manually using the procedure detailed in Supplementary Note 6E.

Hierarchical clustering of mutation signatures. Hierarchical clustering of all 115 samples, by nucleotide mutational context, using the heatmap.2 function from the gplots library (http://cran.r-project.org/web/packages/gplots/index.html) implemented in R 2.15.1 was performed. Mutation counts were scaled within each sample (that is, converted to fraction of mutations corresponding to each category) and clustered using Ward's minimum variance method[45]. Analysis of associations between mutation signature clusters and epidemiological factors (age, histology, geography, tumour grade and smoking status) was performed using Kruskal–Wallis, for continuous factors (age), and Fisher's exact test, for discrete factors (histology, geography, tumour grade and smoking status). A *P*-value threshold of 0.05 was used to decide association for all statistical testing. A detailed description and results can be found in Supplementary Note 8A. For display purposes in Fig. 1, Tp*CpG mutations (which belong to both of the Tp*C and *CpG groups) are redistributed proportionately to each group, on the basis of the relative frequencies of the other Tp*C and*CpG mutations in each tumour.

Hierarchical clustering of copy-number variants. Copy-number profiles, generated by GISTIC 2.0, were clustered into three categories. Each category was tested for association with tumour grade and histology using Fisher's exact test and a *P*-value threshold of 0.05 (Supplementary Note 10B).

Gene expression analysis. Gene expression was measured using Cufflinks and cuffdiff[46] from the RNA-seq data and fragments per kilobase of exon per million fragments mapped (FPKM) was obtained. As non-transcribed genes tend to have more artefactual mutations[37,38,47,48], genes with low expression values (FPKM < 1) were filtered from significantly mutated gene lists produced by MutSig2.0 (Supplementary Note 6C). Genes with detected HPV integration sites were further analysed for increased expression levels by ranking the expression levels of relevant genes for each sample (Supplementary Notes 11E, F and G). Consensus clustering was performed on the RNA-seq-derived gene expression data from 79 tumours, using ConsensusClusterPlus[49]. This analysis was run, using 1,000 resampling iterations and a maximum of 25 clusters, on the 5,000 genes with the largest deviation of FPKM scores across all patients. A final *k* value was chosen based on minimum threshold of change as *k* was increased (Supplementary Note 12). The correlative relationships between copy-number variation and gene expression data were evaluated using Pearson correlation (Supplementary Note 10C).

Fusion analysis in transcriptome. RNA-seq data was analysed for fusion events by identifying inter-chromosomal chimaeric read pairs or exon–exon read pairs separated by at least 1 Mb, with pairmates in the appropriate coding strand orientation. Unmapped reads spanning the putative fusion junction were also identified. High-confidence fusion events are defined as having at least three reads mapped to a junction fusion (Supplementary Note 13B).

HPV typing. HPV typing was done by two multiplex HPV DNA PCR methods: the fluorescent f-HPV assay[50] and the mass-spectrometry-based HPV PCR-MassArray[51,52] (Supplementary Note 3A). In addition, PathSeq[53] was used to generate HPV typing information from RNA-seq data (Supplementary Note 3B).

HPV genome integration analysis. PathSeq[53] was used to identify sites of HPV genome integration in the cohort (Supplementary Notes 7 and 11A). Integration sites were identified using paired reads where one read aligned to the HPV genome (using NCBI viral databases) and the other to the human genome (Supplementary Note 11C). Validation of a recurrent integration site, in *RAD51B*, was performed by RT–PCR using primers targeting the junctions of specific HPV-*RAD51B* chimaeric reads in three tumours (Supplementary Note 11B). In cases in which the tumour with an integration site also had SNP array data, the distance between the integration site and the nearest copy-number amplification was calculated. SNP Array data were used to identify amplifications. To evaluate co-occurrence of integration sites and copy-number amplifications, data for a null model was created by simulating integration sites at uniformly distributed locations in the genome and assigning the simulated integration sites to random samples. The distribution of distances between the simulated integration sites and nearest copy-number amplification compared to the true distribution of distances (Supplementary Note 11D). In Fig. 3a, the true data (top histogram) was compared to distances generated from 100,000 permutations of randomly distributed integration sites across the genome with respect to the observed amplifications (bottom histogram). Overlapping amplification/integration sites have a distance of 0. Integration sites without amplification on the same chromosome were assigned a distance of the longest chromosome plus 1 (the bars on the right). MSigDB[54] was queried to evaluate whether the HPV integration sites co-occur in pathways with known roles in cancer (Supplementary Note 11H).

Epidemiological analysis for mutation rates. The statistical significance of mutation rate across histological types was corrected for the epidemiological factors: age (continuous), geography (discrete), tumour grade (discrete) and smoking status at diagnosis (discrete). The Fisher's exact test (for discrete factors) and the Wilcoxon test (for continuous factors) were run across histology (squamous cell and adenocarcinomas only). A *P*-value threshold of 0.05 was used to decide association for all statistical testing. For factors for which an association was found with histology and mutation rate, we used a linear regression model to test whether histology was an independent predictor of mutation rate (Supplementary Note 8B).

Analysis of miscellaneous genes and pathways of interest. Additional analysis of notable gene sets and pathways were performed using different techniques depending on the task. *APOBEC* gene family members are known to deaminate cytosine residues in the Tp*C context[55,56]. The mutational characteristics of these genes were investigated and results reported in Supplementary Note 14A. The Fanconi anaemia pathway has been implicated in suppressing HPV infection. Mutated genes in the Fanconi anaemia pathway were identified from the somatic single nucleotide variant, somatic indel, and germline calls (Supplementary Note 14B). *TERC* has been identified as a marker for genomic instability and copy number gain. Therefore, we identified copy-number variations encompassing the *TERC*

gene as well as somatic and germline single nucleotide variants and indel mutations (Supplementary Notes 14C).

31. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnol.* **27,** 182–189 (2009).
32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
33. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
34. Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333,** 1157–1160 (2011).
35. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486,** 405–409 (2012).
36. Lee, R. S. *et al.* A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers. *J. Clin. Invest.* **122,** 2983–2988 (2012).
37. Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470,** 214–220 (2011).
38. Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471,** 467–472 (2011).
39. Cibulskis, K. *et al.* ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* **27,** 2601–2602 (2011).
40. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39,** D945–D950 (2011).
41. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43,** 491–498 (2011).
42. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
43. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnol.* **30,** 413–421 (2012).
44. Erlich, H. HLA DNA typing: past, present, and future. *Tissue Antigens* **80,** 1–11 (2012).
45. Ward, J. H., Jr. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58,** 236–244 (1963).
46. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnol.* **31,** 46–63 (2012).
47. Bass, A. J. *et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent *VTI1A-TCF7L2* fusion. *Nature Genet.* **43,** 964–968 (2011).
48. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463,** 191–196 (2010).
49. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26,** 1572–1573 (2010).
50. Cañadas, M. P. *et al.* Comparison of the f-HPV typing and Hybrid Capture II assays for detection of high-risk HPV genotypes in cervical samples. *J. Virol. Methods* **183,** 14–18 (2012).
51. Walline, H. M. *et al.* High-risk human papillomavirus detection in oropharyngeal, nasopharyngeal, and, oral cavity cancers: comparison of multiple methods. *JAMA Otolaryngol. Head Neck Surg.* http://dx.doi.org/10.1001/jamaoto.2013.5460. (31 October 2013).
52. Yang, H. *et al.* Sensitive detection of human papillomavirus in cervical, head/neck, and schistosomiasis-associated bladder malignancies. *Proc. Natl Acad. Sci. USA* **102,** 7683–7688 (2005).
53. Kostic, A. D. *et al.* PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nature Biotechnol.* **29,** 393–396 (2011).
54. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102,** 15545–15550 (2005).
55. Harris, R. S. & Liddament, M. T. Retroviral restriction by APOBEC proteins. *Nature Rev. Immunol.* **4,** 868–877 (2004).
56. Roberts, S. A. *et al.* An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nature Genet.* **45,** 970–976 (2013).

LETTER

doi:10.1038/nature12873

Genetics of rheumatoid arthritis contributes to biology and drug discovery

A list of authors and their affiliations appears at the end of the paper

A major challenge in human genetics is to devise a systematic strategy to integrate disease-associated variants with diverse genomic and biological data sets to provide insight into disease pathogenesis and guide drug discovery for complex traits such as rheumatoid arthritis (RA)[1]. Here we performed a genome-wide association study meta-analysis in a total of >100,000 subjects of European and Asian ancestries (29,880 RA cases and 73,758 controls), by evaluating ~10 million single-nucleotide polymorphisms. We discovered 42 novel RA risk loci at a genome-wide level of significance, bringing the total to 101 (refs 2–4). We devised an *in silico* pipeline using established bioinformatics methods based on functional annotation[5], *cis*-acting expression quantitative trait loci[6] and pathway analyses[7–9]—as well as novel methods based on genetic overlap with human primary immunodeficiency, haematological cancer somatic mutations and knockout mouse phenotypes—to identify 98 biological candidate genes at these 101 risk loci. We demonstrate that these genes are the targets of approved therapies for RA, and further suggest that drugs approved for other indications may be repurposed for the treatment of RA. Together, this comprehensive genetic study sheds light on fundamental genes, pathways and cell types that contribute to RA pathogenesis, and provides empirical evidence that the genetics of RA can provide important information for drug discovery.

We conducted a three-stage trans-ethnic meta-analysis (Extended Data Fig. 1). On the basis of the polygenic architecture of RA[10] and shared genetic risk among different ancestry[3,4], we proposed that combining a genome-wide association study (GWAS) of European and Asian ancestry would increase power to detect novel risk loci. In stage 1, we combined 22 GWAS for 19,234 cases and 61,565 controls of European and Asian ancestry[2–4]. We performed trans-ethnic, European-specific and Asian-specific GWAS meta-analysis by evaluating ~10 million single-nucleotide polymorphisms (SNPs)[11]. Characteristics of the cohorts, genotyping platforms and quality control criteria are described in Extended Data Table 1 (overall genomic control inflation factor $\lambda_{GC} < 1.075$).

Stage 1 meta-analysis identified 57 loci that satisfied a genome-wide significance threshold of $P < 5.0 \times 10^{-8}$, including 17 novel loci (Extended Data Fig. 2). We then conducted a two-step replication study (stage 2 for *in silico* and stage 3 for *de novo*) in 10,646 RA cases and 12,193 controls for the loci with $P < 5.0 \times 10^{-6}$ in stage 1. In a combined analysis of stages 1–3, we identified 42 novel loci with $P < 5.0 \times 10^{-8}$ in any of the trans-ethnic, European or Asian meta-analyses. This increases the total number of RA risk loci to 101 (Table 1 and Supplementary Table 1).

Comparison of 101 RA risk loci revealed significant correlations of risk allele frequencies (RAFs) and odds ratios (ORs) between Europeans and Asians (Extended Data Fig. 3a–c; Spearman's $\rho = 0.67$ for RAF and 0.76 for OR; $P < 1.0 \times 10^{-13}$), although five loci demonstrated population-specific associations ($P < 5.0 \times 10^{-8}$ in one population but $P > 0.05$ in the other population without overlap of the 95% confidence intervals (95% CIs) of the ORs). In the population-specific genetic risk model, the 100 RA risk loci outside of the major histocompatibility complex (MHC) region[12] explained 5.5% and 4.7% of heritability in Europeans and Asians, respectively, with 1.6% of the heritability explained by the novel loci. The trans-ethnic genetic risk model, based on the RAF from one population but the OR from the other population, could explain the majority (>80%) of the known heritability in each population (4.7% for Europeans and 3.8% for Asians). These observations support our hypothesis that the genetic risk of RA is shared, in general, among Asians and Europeans.

We assessed enrichment of 100 non-MHC RA risk loci in epigenetic chromatin marks[13] (Extended Data Fig. 3d). Of 34 cell types investigated, we observed significant enrichment of RA risk alleles with trimethylation of histone H3 at lysine 4 (H3K4me3) peaks in primary $CD4^+$ regulatory T cells (T_{reg} cells; $P < 1.0 \times 10^{-5}$). For the RA risk loci enriched with T_{reg} H3K4me3 peaks, we incorporated the epigenetic annotations along with trans-ethnic differences in patterns of linkage disequilibrium to fine-map putative causal risk alleles (Extended Data Fig. 3e, f).

We found that approximately two-thirds of RA risk loci demonstrated pleiotropy with other human phenotypes (Extended Data Fig. 4), including immune-related diseases (for example, vitiligo, primary biliary cirrhosis), inflammation-related or haematological biomarkers (for example, fibrinogen, neutrophil counts) and other complex traits (for example, cardiovascular diseases).

Each of 100 non-MHC RA risk loci contains on average ~4 genes in the region of linkage disequilibrium (in total 377 genes). To prioritize systematically the most likely biological candidate gene, we devised an *in silico* bioinformatics pipeline. In addition to the published methods that integrate data across associated loci[7,8], we evaluated several biological data sets to test for enrichment of RA risk genes, which helps to pinpoint a specific gene in each loci (Extended Data Figs 5, 6 and Supplementary Tables 2–4).

We first conducted functional annotation of RA risk SNPs. Sixteen per cent of SNPs were in linkage disequilibrium with missense SNPs ($r^2 > 0.80$; Extended Data Fig. 5a, b). The proportion of missense RA risk SNPs was higher compared with a set of genome-wide common SNPs (8.0%), and relatively much higher in the explained heritability (~26.8%). Using *cis*-acting expression quantitative trait loci (*cis*-eQTL) data obtained from peripheral blood mononuclear cells (5,311 individuals)[6] and from $CD4^+$ T cells and $CD14^+CD16^-$ monocytes (212 individuals), we found that RA risk SNPs in 44 loci showed *cis*-eQTL effects (false discovery rate (FDR) q or permutation $P < 0.05$; Extended Data Table 2).

Second, we evaluated whether genes from RA risk loci overlapped with human primary immunodeficiency (PID) genes[14], and observed significant overlap (14/194 = 7.2%, $P = 1.2 \times 10^{-4}$; Fig. 1a and Extended Data Fig. 5c). Classification categories of PID genes showed different patterns of overlap: the highest proportion of overlap was in 'immune dysregulation' (4/21 = 19.0%, $P = 0.0033$) but there was no overlap in 'innate immunity'.

Third, we evaluated overlap with cancer somatic mutation genes[15], under the hypothesis that genes with cell growth advantages may contribute to RA development. Among 444 genes with registered cancer somatic mutations[15], we observed significant overlap with genes implicated in haematological cancers (17/251 = 6.8%, $P = 1.2 \times 10^{-4}$; Fig. 1b and Extended Data Fig. 5d), but not with genes implicated in non-haematological cancers (6/221 = 2.7%, $P = 0.56$).

Table 1 | Novel rheumatoid arthritis risk loci identified by trans-ethnic GWAS meta-analysis in >100,000 subjects

SNP	Chr	Genes	A1/A2 (+)	Trans-ethnic		European		Asian	
				OR (95% CI)	*P*	OR (95% CI)	*P*	OR (95% CI)	*P*
rs227163	1	*TNFRSF9*	C/T	1.04 (1.02–1.06)	3.9×10^{-4}	1.00 (0.97–1.03)	9.3×10^{-1}	1.11 (1.08–1.16)*	3.1×10^{-9}*
rs28411352	1	*MTF1-INPP5B*	T/C	1.11 (1.08–1.14)*	2.8×10^{-12}*	1.10 (1.07–1.14)*	5.9×10^{-9}*	1.12 (1.06–1.19)	7.8×10^{-5}
rs2105325	1	*LOC100506023*	C/A	1.12 (1.08–1.15)*	6.9×10^{-13}*	1.12 (1.08–1.15)*	3.3×10^{-11}*	1.13 (1.04–1.23)	5.2×10^{-3}
rs10175798	2	*LBH*	A/G	1.08 (1.06–1.11)*	1.1×10^{-9}*	1.09 (1.06–1.12)*	4.2×10^{-8}*	1.07 (1.02–1.13)	6.4×10^{-3}
rs6732565	2	*ACOXL*	A/G	1.07 (1.05–1.10)*	2.7×10^{-8}*	1.10 (1.07–1.14)*	9.4×10^{-9}*	1.04 (1.00–1.08)	4.0×10^{-2}
rs6715284	2	*CFLAR-CASP8*	G/C	1.15 (1.10–1.20)*	1.8×10^{-9}*	1.15 (1.10–1.20)*	2.5×10^{-9}*	-	-
rs4452313	3	*PLCL2*	T/A	1.09 (1.06–1.12)*	1.6×10^{-10}*	1.11 (1.08–1.15)*	5.2×10^{-11}*	1.04 (0.99–1.09)	9.2×10^{-2}
rs3806624	3	*EOMES*	G/A	1.08 (1.05–1.11)*	8.6×10^{-9}*	1.08 (1.05–1.12)*	2.8×10^{-8}*	1.06 (0.99–1.14)	1.0×10^{-1}
rs9826828	3	*IL20RB*	A/G	1.44 (1.28–1.61)*	8.6×10^{-10}*	1.44 (1.28–1.61)*	8.7×10^{-10}*	-	-
rs13142500	4	*CLNK*	C/T	1.10 (1.07–1.13)*	3.0×10^{-9}*	1.10 (1.06–1.15)	2.4×10^{-6}	1.10 (1.04–1.15)	2.8×10^{-4}
rs2664035	4	*TEC*	A/G	1.07 (1.04–1.10)	9.5×10^{-8}	1.08 (1.05–1.11)*	3.3×10^{-8}*	1.03 (0.97–1.08)	3.3×10^{-1}
rs9378815	6	*IRF4*	C/G	1.09 (1.06–1.12)*	1.7×10^{-10}*	1.09 (1.05–1.12)	1.4×10^{-7}	1.10 (1.04–1.15)	2.3×10^{-4}
rs2234067	6	*ETV7*	C/A	1.15 (1.10–1.20)*	1.6×10^{-9}*	1.14 (1.09–1.19)*	4.1×10^{-8}*	1.22 (1.06–1.41)	7.0×10^{-3}
rs9373594	6	*PPIL4*	T/C	1.09 (1.06–1.12)*	3.0×10^{-9}*	1.07 (1.02–1.12)	6.5×10^{-3}	1.11 (1.07–1.15)*	4.8×10^{-8}*
rs67250450	7	*JAZF1*	T/C	1.10 (1.07–1.14)*	3.7×10^{-9}*	1.11 (1.07–1.14)*	2.6×10^{-9}*	1.02 (0.84–1.23)	8.5×10^{-1}
rs4272	7	*CDK6*	G/A	1.10 (1.06–1.13)*	5.0×10^{-9}*	1.10 (1.07–1.14)*	1.2×10^{-8}*	1.06 (0.98–1.15)	1.3×10^{-1}
rs998731	8	*TPD52*	T/C	1.08 (1.05–1.11)*	1.9×10^{-8}*	1.09 (1.06–1.12)*	6.6×10^{-9}*	1.02 (0.96–1.10)	4.9×10^{-1}
rs678347	8	*GRHL2*	G/A	1.08 (1.05–1.11)*	1.6×10^{-8}*	1.10 (1.06–1.13)*	7.3×10^{-9}*	1.03 (0.98–1.10)	2.6×10^{-1}
rs1516971	8	*PVT1*	T/C	1.15 (1.10–1.20)*	1.3×10^{-10}*	1.16 (1.11–1.21)*	3.2×10^{-11}*	-	-
rs12413578	10	10p14	C/T	1.20 (1.13–1.29)*	4.8×10^{-8}*	1.20 (1.12–1.29)	7.5×10^{-8}	-	-
rs793108	10	*ZNF438*	T/C	1.08 (1.05–1.10)*	1.3×10^{-9}*	1.07 (1.04–1.10)	6.1×10^{-7}	1.09 (1.04–1.14)	4.4×10^{-4}
rs2671692	10	*WDFY4*	A/G	1.07 (1.05–1.10)*	2.8×10^{-9}*	1.06 (1.03–1.09)	2.6×10^{-5}	1.10 (1.05–1.14)	9.9×10^{-6}
rs726288	10	*SFTPD*	T/C	1.14 (1.07–1.20)	1.6×10^{-5}	0.96 (0.86–1.06)	4.1×10^{-1}	1.22 (1.14–1.31)*	8.8×10^{-9}*
rs968567	11	*FADS1-FADS2-FADS3*	C/T	1.12 (1.07–1.16)*	1.8×10^{-8}*	1.12 (1.07–1.16)*	1.8×10^{-8}*	-	-
rs4409785	11	*CEP57*	C/T	1.12 (1.09–1.16)*	1.2×10^{-11}*	1.12 (1.08–1.16)*	3.6×10^{-9}*	1.16 (1.07–1.27)	4.3×10^{-4}
chr11:107967350	11	*ATM*	A/G	1.21 (1.13–1.29)*	1.4×10^{-8}*	1.21 (1.13–1.29)*	1.1×10^{-8}*	-	-
rs73013527	11	*ETS1*	C/T	1.09 (1.06–1.12)*	1.2×10^{-10}*	1.08 (1.05–1.11)	1.0×10^{-6}	1.14 (1.08–1.21)	4.1×10^{-6}
rs773125	12	*CDK2*	A/G	1.09 (1.06–1.12)*	1.1×10^{-10}*	1.09 (1.06–1.12)*	2.1×10^{-8}*	1.10 (1.04–1.17)	1.1×10^{-3}
rs10774624	12	*SH2B3-PTPN11*	G/A	1.09 (1.06–1.13)*	6.8×10^{-9}*	1.09 (1.06–1.13)*	6.9×10^{-9}*	-	-
rs9603616	13	*COG6*	C/T	1.10 (1.07–1.13)*	1.6×10^{-12}*	1.11 (1.07–1.14)*	2.8×10^{-11}*	1.08 (1.02–1.14)	1.0×10^{-2}
rs3783782	14	*PRKCH*	A/G	1.14 (1.09–1.18)*	2.2×10^{-9}*	1.12 (0.96–1.31)	1.4×10^{-1}	1.14 (1.09–1.19)*	4.4×10^{-9}*
rs1950897	14	*RAD51B*	T/C	1.10 (1.07–1.13)*	8.2×10^{-11}*	1.09 (1.06–1.12)*	5.0×10^{-8}*	1.16 (1.08–1.25)	1.1×10^{-4}
rs4780401	16	*TXNDC11*	T/G	1.07 (1.05–1.10)*	4.1×10^{-8}*	1.09 (1.06–1.13)*	8.7×10^{-9}*	1.03 (0.98–1.08)	2.5×10^{-1}
rs72634030	17	*C1QBP*	A/C	1.12 (1.08–1.17)*	1.5×10^{-9}*	1.12 (1.06–1.19)	2.9×10^{-5}	1.12 (1.07–1.18)	9.6×10^{-6}
rs1877030	17	*MED1*	C/T	1.09 (1.06–1.12)*	1.9×10^{-8}*	1.09 (1.05–1.13)	1.3×10^{-5}	1.09 (1.04–1.14)	3.2×10^{-4}
rs2469434	18	*CD226*	C/T	1.07 (1.05–1.10)*	8.9×10^{-10}*	1.05 (1.02–1.08)	6.7×10^{-4}	1.11 (1.07–1.15)*	1.2×10^{-8}*
chr19:10771941	19	*ILF3*	C/T	1.47 (1.30–1.67)*	8.6×10^{-10}*	1.47 (1.30–1.67)*	8.8×10^{-10}*	-	-
rs73194058	21	*IFNGR2*	C/A	1.08 (1.05–1.12)	1.2×10^{-6}	1.13 (1.08–1.18)*	2.6×10^{-8}*	1.03 (0.98–1.08)	2.9×10^{-1}
rs1893592	21	*UBASH3A*	A/C	1.11 (1.08–1.14)*	7.2×10^{-12}*	1.11 (1.07–1.15)*	9.8×10^{-9}*	1.11 (1.05–1.18)	1.3×10^{-4}
rs11089637	22	*UBE2L3-YDJC*	C/T	1.08 (1.05–1.11)*	2.1×10^{-9}*	1.10 (1.06–1.15)	2.0×10^{-7}	1.06 (1.02–1.10)	8.9×10^{-4}
rs909685	22	*SYNGR1*	A/T	1.13 (1.10–1.16)*	1.4×10^{-16}*	1.11 (1.08–1.15)*	6.4×10^{-12}*	1.23 (1.14–1.33)	2.0×10^{-7}
chrX:78464616	X	*P2RY10*	A/C	1.11 (1.07–1.15)*	3.5×10^{-8}*	1.16 (0.78–1.75)	4.6×10^{-1}	1.11 (1.07–1.15)*	3.6×10^{-8}*

SNPs newly associated with $P < 5.0 \times 10^{-8}$ in the combined study of the stage 1 GWAS meta-analysis and the stages 2 and 3 replication studies of trans-ethnic (Europeans and Asians), European or Asian ancestry are indicated. SNPs, positions and alleles are based on the positive (+) strand of NCBI build 37. A1 represents an RA risk allele. Chr, chromosome; OR, odds ratio; 95% CI, 95% confidence interval. Full results of the studies are available in Supplementary Table 1. Hyphens between gene names indicate that several candidate RA risk genes were included in the region.
*Association results with $P < 5.0 \times 10^{-8}$.

Fourth, we evaluated overlap with genes implicated in knockout mouse phenotypes[16]. Among the 30 categories of phenotypes[16], we observed 3 categories significantly enriched with RA risk genes ($P < 0.05/30 = 0.0017$): 'haematopoietic system phenotype', 'immune system phenotype', and 'cellular phenotype' (Extended Data Fig. 5e).

Last, we conducted molecular pathway enrichment analysis (Fig. 1c and Extended Data Fig. 5f). We observed enrichment (FDR $q < 0.05$) for T-cell-related pathways, consistent with cell-specific epigenetic marks, as well as enrichment for B-cell and cytokine signalling pathways (for example, interleukin (IL)-10, interferon, granulocyte–macrophage colony-stimulating factor (GM-CSF)). For comparison, our previous RA GWAS meta-analysis[2] did not identify the B-cell and cytokine signalling pathways, thereby indicating that as more loci are discovered, further biological pathways are identified.

On the basis of these new findings, we adopted the following 8 criteria to prioritize each of the 377 genes from the 100 non-MHC RA risk loci (Fig. 2 and Extended Data Fig. 6a–c): (1) genes with RA risk missense variant ($n = 19$); (2) *cis*-eQTL genes ($n = 51$); (3) genes prioritized by PubMed text mining[7] ($n = 90$); (4) genes prioritized by protein–protein interaction (PPI)[8] ($n = 63$); (5) PID genes ($n = 15$); (6) haematological cancer somatic mutation genes ($n = 17$); (7) genes prioritized by associated knockout mouse phenotypes ($n = 86$); and (8) genes prioritized by molecular pathway analysis[9] ($n = 35$).

Ninety-eight genes (26.0%) had a score ≥2, which we defined as 'candidate biological RA risk genes'. Nineteen loci included multiple biological RA risk genes (for example, *IL3* and *CSF2* at chromosome 5q31), whereas no biological gene was selected from 40 loci (Supplementary Table 5).

To provide empirical evidence of the pipeline, we evaluated relationships of the gene scores to independent genomic or epigenetic information. Genes with higher biological scores were more likely to be the nearest gene to the risk SNP (18.6% for gene score <2 and 49.0% for gene score ≥2; $P = 2.1 \times 10^{-8}$), and also to be included in the region where RA risk SNPs were overlapping with H3K4me3 T_{reg} peaks (41.9% for gene score <2 and 57.1% for gene score ≥2; $P = 0.034$). Further, T_{reg} cells demonstrated the largest increase in overlapping proportions with H3K4me3 peaks for increase of biological gene scores compared with other cell types (Extended Data Fig. 6d).

Finally, we evaluated the potential role of RA genetics in drug discovery. We proposed that if human genetics is useful for drug target validation, then it should identify existing approved drugs for RA. To test this 'therapeutic hypothesis'[1], we obtained 871 drug target genes corresponding to approved, in clinical trials or experimental drugs for human diseases[17,18] (Supplementary Table 6). We evaluated whether any of the protein products from the identified biological RA risk genes, or any genes from a direct PPI network with such protein products

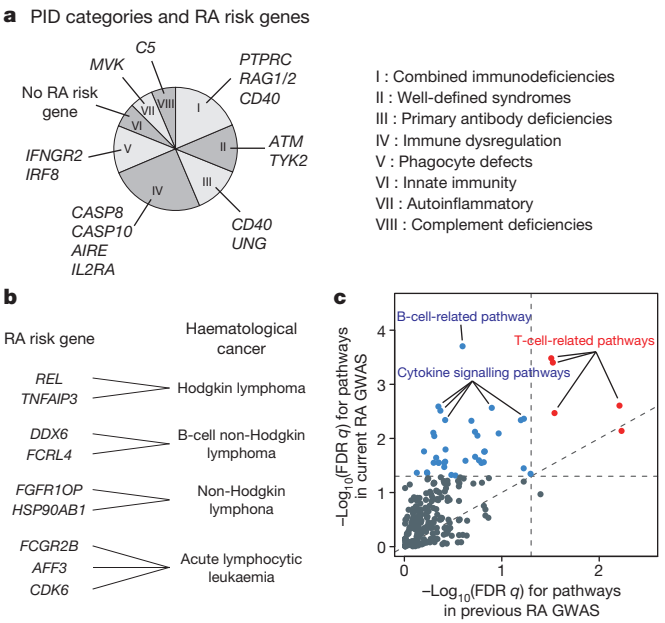


Figure 1 | Overlap of RA risk loci with PID genes, haematological cancer somatic mutations and molecular pathways. **a**, Overlap of RA risk genes with PID genes, subdivided by PID categories (I–VIII). **b**, Examples of overlap of haematological cancer somatic mutation genes with RA risk genes. **c**, Comparisons of molecular pathway analysis results between the current trans-ethnic meta-analysis (*y*-axis) and the previous meta-analysis for RA (*x*-axis)[2]. Each dot represents a molecular pathway. Dotted line represents FDR $q = 0.05$ or $y = x$.

(Fig. 3a), are the pharmacologically active targets of approved RA drugs (Extended Data Fig. 7a).

Twenty-seven drug target genes of approved RA drugs demonstrated significant overlap with 98 biological RA risk genes and 2,332 genes from the expanded PPI network (18 genes overlapped; 3.7-fold enrichment by permutation analysis, $P < 1.0 \times 10^{-5}$; Fig. 3b). For comparison, all drug target genes (regardless of disease indication) overlapped with 247 genes, which is 1.7-fold more enrichment than expected by chance, but less than 2.2-fold enrichment compared with overlap of the target genes of RA drugs ($P = 0.0035$). Examples of approved RA therapies identified by this analysis include tocilizumab[19,20] (anti-IL6R), tofacitinib[21] (JAK3 inhibitor) and abatacept[21] (CTLA4–immunoglobulin; Fig. 3c and Extended Data Fig. 8).

We also assessed how approved drugs for other diseases might be connected to biological RA risk genes. We highlight *CDK6* and *CDK4*, targets of three approved drugs for different types of cancer[22] (Fig. 3d).

RA risk SNP (cytoband)	Gene	Score	Biological gene criteria: RA risk missense variant	*cis*-eQTL	PubMed text mining	PPI	PID	Haematological cancer	Knockout mouse phenotype	Molecular pathway	Nearest gene from RA risk SNP	Overlap with H3K4me3 peaks: T_{reg} primary cells	CD4+ memory primary cells	CD4+ naive primary cells	CD8+ memory primary cells	CD8+ naive primary cells	CD34+ primary cells	CD34+ cultured cells	Mobilized CD34+ primary cells	CD19+ primary cells	CD3+ primary cells	Drug target gene	RA drug target gene	PPI with RA drug target gene
chr1:2523811 (1p36)	*TNFRSF14*	4			■				■	■		■	■	■	■	■	■	■	■	■	■			■
rs2301888 (1p36)	*PADI4*	2	■		■						■													
rs2476601 (1p13)	*PTPN22*	5	■	■	■	■			■		■								■					
rs2228145 (1q21)	*IL6R*	5	■	■	■					■	■	■		■	■	■					■	■	■	■
chr1:161644258 (1q23)	*FCGR2B*	5		■	■			■	■	■	■											■		
rs17668708 (1q31)	*PTPRC*	6		■	■	■	■		■	■	■	■	■	■	■	■		■	■	■	■			■
rs34695944 (2p16-p15)	*REL*	4			■	■		■	■		■	■	■	■	■	■	■	■	■	■	■			
rs9653442 (2q11)	*AFF3*	4		■	■	■		■			■													
rs11889341 (2q32)	*STAT4*	3			■				■	■	■	■		■	■									■
rs6715284 (2q33)	*CFLAR*	3			■				■	■	■													■
rs1980422 (2q33)	*CD28*	4		■	■				■	■	■	■	■	■	■	■			■		■	■		■
rs3087243 (2q33)	*CTLA4*	4			■	■			■	■	■	■										■		■
rs45475795 (4q26-q27)	*IL2*	5			■	■		■	■	■	■	■	■	■	■	■			■					
rs657075 (5q31)	*IL3*	4			■	■			■	■														■
rs657075 (5q31)	*CSF2*	4			■	■			■	■	■													■
rs2233424 (6p21)	*NFKBIE*	4	■		■	■			■	■	■	■	■	■	■	■	■	■	■	■	■			■
rs7752903 (6q23)	*TNFAIP3*	6	■		■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■			■
rs1571878 (6q27)	*CCR6*	2			■						■	■	■	■	■					■				
rs4272 (7q21)	*CDK6*	4			■			■	■	■	■	■										■		
chr7:128580042 (7q32)	*IRF5*	4		■	■	■			■		■		■	■	■		■	■	■	■	■			
rs10985070 (9q33)	*TRAF1*	4		■	■	■			■				■	■	■	■	■	■	■	■	■			
rs10985070 (9q33)	*C5*	4		■	■		■		■				■	■	■	■	■	■	■	■	■	■		
rs706778 (10p15)	*IL2RA*	5		■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■		■
rs331463 (11p12)	*TRAF6*	4			■	■			■	■	■		■	■	■	■		■	■	■	■			
rs331463 (11p12)	*RAG1*	4			■	■	■		■				■	■	■	■	■	■	■	■	■			
rs508970 (11q12)	*CD5*	4	■		■	■			■				■	■	■	■	■		■		■			
chr11:107967350 (11q22)	*ATM*	4					■	■	■	■														■
rs773125 (12q13)	*CDK2*	3				■			■	■			■	■	■	■	■	■	■	■	■	■		
rs1633360 (12q13-q14)	*CDK4*	3			■	■			■	■			■	■	■	■	■	■	■	■	■	■		
rs10774624 (12q24)	*SH2B3*	5	■	■	■	■			■	■	■			■	■	■			■					
chr17:38031857 (17q12-q21)	*IKZF3*	4		■	■	■			■	■			■	■	■	■	■	■	■	■	■			
chr17:38031857 (17q12-q21)	*CSF3*	4			■	■			■	■			■	■	■	■	■	■	■	■	■			■
rs8083786 (18p11)	*PTPN2*	3			■	■			■	■	■		■	■						■	■			■
rs34536443 (19p13)	*ICAM1*	4			■	■			■	■												■		
rs34536443 (19p13)	*TYK2*	6	■		■	■	■		■	■														■
rs4239702 (20q13)	*CD40*	6		■	■	■	■		■	■	■		■	■	■	■	■	■	■	■	■	■		■
rs73194058 (21q22)	*IFNGR2*	6		■	■	■	■		■	■	■		■	■	■	■	■	■	■	■	■	■		■
rs2236668 (21q22)	*ICOSLG*	5	■	■	■				■	■	■													
rs2236668 (21q22)	*AIRE*	4			■	■	■																	
rs3218251 (22q12)	*IL2RB*	3			■	■			■													■		■
rs5987194 (Xq28)	*IRAK1*	3	■		■	■			■				■	■	■	■	■	■	■	■	■			■

Figure 2 | Prioritized biological RA risk genes. Representative biological RA risk genes. We list the summary gene score derived from individual criteria (filled red box indicates criterion satisfied; 98 genes with a score ≥2 out of 377 genes included in the RA risk loci were defined as 'biological candidate genes'; see Extended Data Fig. 6). Filled blue boxes indicate the nearest gene to the RA risk SNP. Filled green boxes indicate overlap with H3K4me3 peaks in immune-related cells. Filled purple boxes indicate overlap with drug target genes. For full results, see Supplementary Table 5.

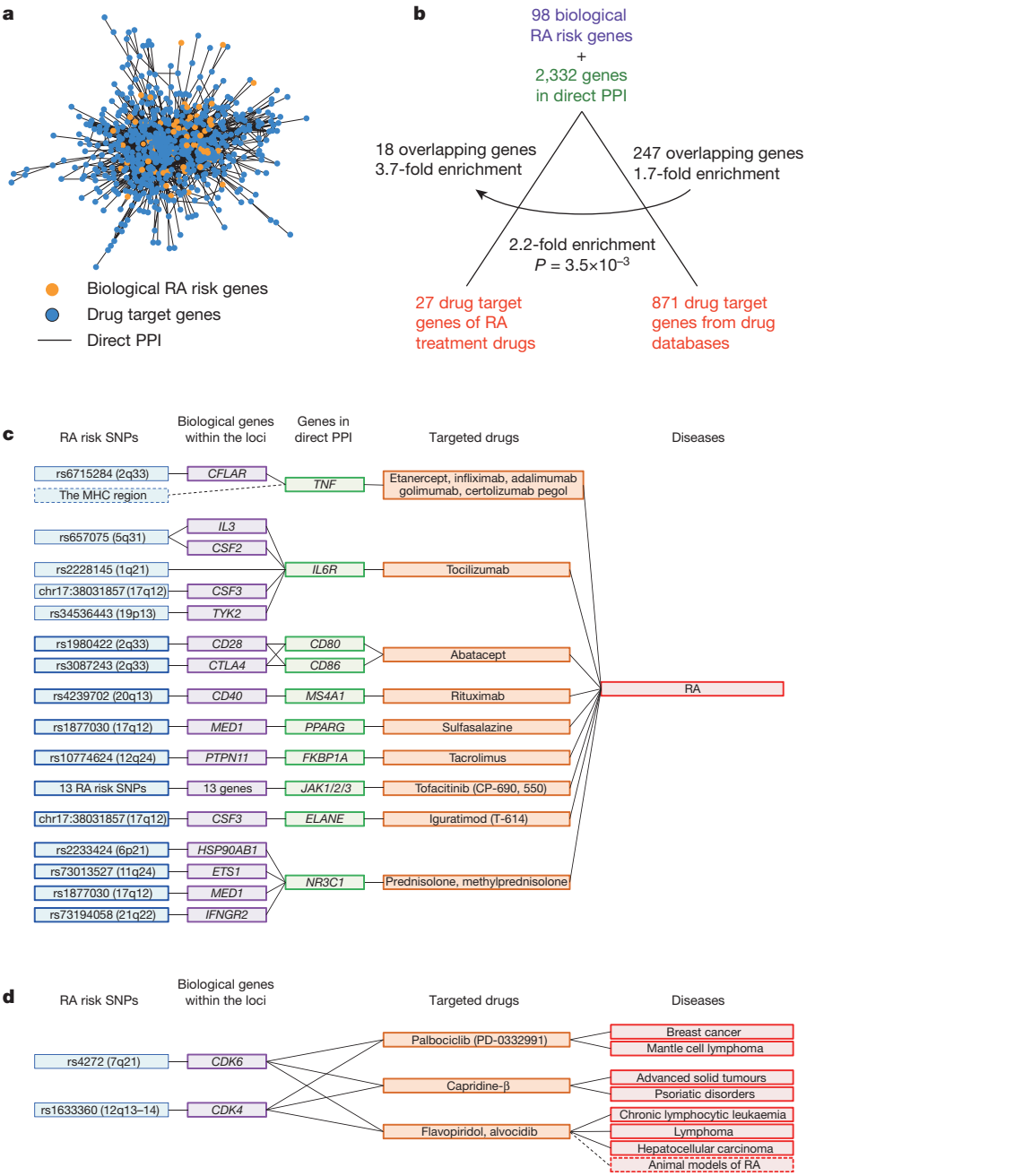


Figure 3 | Connection of biological RA risk genes to drug targets. **a**, PPI network of biological RA risk genes and drug target genes. **b**, Overlap and relative enrichment of 98 biological RA risk genes with targets of approved RA drugs and with all drug target genes. Enrichment was more apparent than that from all 377 RA risk genes (Extended Data Fig. 7c). **c**, Connections between RA risk SNPs (blue), biological genes (purple), genes from PPI (green) and approved RA drugs (orange). For full results, see Extended Data Fig. 8. **d**, Connections between RA genes and drugs indicated for other diseases.

In support for repurposing, one *CDK6/CDK4* inhibitor, flavopiridol, has been shown to ameliorate disease activity in animal models of RA[22]. Further, the biology is plausible, as several approved RA drugs were initially developed for cancer treatment and then repurposed for RA (for example, rituximab). Although further investigations are necessary, we propose that target genes/drugs selected by this approach could represent promising candidates for novel drug discovery for RA treatment.

We note that a non-random distribution of drug-to-disease indications in the databases could potentially bias our results. Namely, because RA risk genes are enriched for genes with immune function, spurious enrichment with drug targets could occur if the majority of drug indications in databases were for immune-mediated diseases or immune-related target genes. However, such enrichment was not evident in our analysis (~11% for drug indications and ~9% for target genes; Extended Data Fig. 7b).

Through a comprehensive genetic study with >100,000 subjects, we identified 42 novel RA risk loci and provided novel insight into RA pathogenesis. We particularly highlight the role of genetics for drug discovery. Although there have been anecdotal examples of this[1,23], our study provides a systematic approach by which human genetic data can be efficiently integrated with other biological information to derive biological insights and drive drug discovery.

METHODS SUMMARY

Details can be found in Methods, Extended Data Fig. 1, Extended Data Table 1 and Supplementary Information, including (1) information about the patient collections;

(2) genotyping, quality control and genotype imputation of GWAS data; (3) genome-wide meta-analysis (stage 1); (4) *in silico* and *de novo* replication studies (stages 2 and 3); (5) trans-ethnic and functional annotations of RA risk SNPs; (6) prioritization of biological candidate genes; and (7) drug target gene enrichment analysis.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 June; accepted 7 November 2013.

Published online 25 December 2013.

1. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nature Rev. Drug Discov.* **12,** 581–594 (2013).
2. Stahl, E. A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genet.* **42,** 508–514 (2010).
3. Okada, Y. *et al.* Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. *Nature Genet.* **44,** 511–516 (2012).
4. Eyre, S. *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nature Genet.* **44,** 1336–1340 (2012).
5. Ferreira, R. C. *et al.* Functional *IL6R* 358Ala allele impairs classical IL-6 receptor signaling and influences risk of diverse inflammatory diseases. *PLoS Genet.* **9,** e1003444 (2013).
6. Westra, H. J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genet.* **45,** 1238–1243 (2013).
7. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5,** e1000534 (2009).
8. Rossin, E. J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.* **7,** e1001273 (2011).
9. Segrè, A. V., Groop, L., Mootha, V. K., Daly, M. J. & Altshuler, D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* **6,** e1001058 (2010).
10. Stahl, E. A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature Genet.* **44,** 483–489 (2012).
11. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
12. Raychaudhuri, S. *et al.* Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nature Genet.* **44,** 291–296 (2012).
13. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genet.* **45,** 124–130 (2013).
14. Parvaneh, N., Casanova, J. L., Notarangelo, L. D. & Conley, M. E. Primary immunodeficiencies: a rapidly evolving story. *J. Allergy Clin. Immunol.* **131,** 314–323 (2013).
15. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39,** D945–D950 (2011).
16. Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A. & Richardson, J. E. The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Res.* **40,** D881–D886 (2012).
17. Knox, C. *et al.* DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **39,** D1035–D1041 (2011).
18. Zhu, F. *et al.* Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res.* **40,** D1128–D1136 (2012).
19. Smolen, J. S. *et al.* Consensus statement on blocking the effects of interleukin-6 and in particular by interleukin-6 receptor inhibition in rheumatoid arthritis and other inflammatory conditions. *Ann. Rheum. Dis.* **72,** 482–492 (2013).
20. Nishimoto, N. *et al.* Study of active controlled tocilizumab monotherapy for rheumatoid arthritis patients with an inadequate response to methotrexate (SATORI): significant reduction in disease activity and serum vascular endothelial growth factor by IL-6 receptor inhibition therapy. *Mod. Rheumatol.* **19,** 12–19 (2009).
21. McInnes, I. B. & Schett, G. The pathogenesis of rheumatoid arthritis. *N. Engl. J. Med.* **365,** 2205–2219 (2011).
22. Sekine, C. *et al.* Successful treatment of animal models of rheumatoid arthritis with small-molecule cyclin-dependent kinase inhibitors. *J. Immunol.* **180,** 1954–1961 (2008).
23. Sanseau, P. *et al.* Use of genome-wide association studies for drug repositioning. *Nature Biotechnol.* **30,** 317–320 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements R.M.P. is supported by National Institutes of Health (NIH) grants R01-AR057108, R01-AR056768, U01-GM092691 and R01-AR059648, and holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund. Y.O. is supported by a grant from the Japan Society of the Promotion of Science. D.W. is supported by a grant from the Australian National Health and Medical Research Council (1036541). G.T. is supported by the Rubicon grant from the Netherlands Organization for Scientific Research. A.Z. is supported by a grant from the Dutch Reumafonds (11-1-101) and from the Rosalind Franklin Fellowship, University of Groningen. S.-C.B., S.-Y.B. and H.-S.L. are supported by the Korea Healthcare technology R&D project, Ministry for Health and Welfare (A121983). J.M., M.A.G.-G. and L.R.-R. are funded by the RETICS program, RIER, RD12/0009 from the Instituto de Salud Carlos III, Health Ministry. S.R.-D. and L.Ä.'s work is supported by the Medical Biobank of Northern Sweden. H.K.C. is supported by NIH (NIAMS) grants R01-AR056291, R01-AR065944, R01-AR056768, P60 AR047785 and R21 AR056042. L.P. and L.K. are supported by a senior investigator grant from the European Research Council. S.R. is supported by NIH grants R01AR063759-01A1 and K08-KAR055688A. P.M.V. is a National Health and Medical Research Council Senior Principal Research Fellow. M.A.B. is funded by the National Health and Medical Research Foundation Senior Principal Research Fellowship, and a Queensland State Government Premier's Fellowship. H.X. is funded by the China Ministry of Science and Technology (973 program grant 2011CB946100), the National Natural Science Foundation of China (grants 30972339, 81020108029 and 81273283), and the Science and Technology Commission of Shanghai Municipality (grants 08XD1400400, 11410701600 and 10JC1418400). K.A.S. is supported by a Canada Research Chair, The Sherman Family Chair in Genomics Medicine, Canadian Institutes for Health Research grant 79321 and Ontario Research Fund grant 05-075. S.M. is supported by Health and Labour Sciences Research Grants. The BioBank Japan Project is supported by the Ministry of Education, Culture, Sports, Science and Technology of the Japanese government. This study is supported by the BE THE CURE (BTCure) project. We thank K. Akari, K. Tokunaga and N. Nishida for supporting the study.

Author Contributions Y.O. carried out the primary data analyses. D.W. managed drug target gene data. G.T. conducted histone mark analysis. T.R., H.-J.W., T.E., A.M., B.E.S., P.L.D. and L.F. conducted eQTL analysis. C.T., K.I., Y.K., K.O., A.S., S.Y., G.X., E.K. and K.A.S. conducted the *de novo* replication study. R.R.G., A.M., W.O., T.B., T.W.B., L.J., J. Yin, L.Y., D.-F.S., J. Yang, P.M.V., M.A.B. and H.X. conducted the *in silico* replication study. E.A.S., D.D., J.C., T.K., R.Y. and A.T. managed GWAS data. All other authors, as well as the members of the RACI and GARNET consortia, contributed to additional analyses and genotype and clinical data enrolments. Y.O. and R.M.P. designed the study and wrote the manuscript, with contributions from all authors on the final version of the manuscript.

Author Information Summary statistics from the GWAS meta-analysis, source codes, and data sources used in this study are available at http://plaza.umin.ac.jp/~yokada/datasource/software.htm. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.M.P. (robert.plenge@merck.com) or Y.O. (yokada.brc@tmd.ac.jp).

Yukinori Okada[1,2,3], Di Wu[1,2,3,4,5], Gosia Trynka[1,2,3], Towfique Raj[2,3,6], Chikashi Terao[7,8], Katsunori Ikari[9], Yuta Kochi[10], Koichiro Ohmura[8], Akari Suzuki[10], Shinji Yoshida[9], Robert R. Graham[11], Arun Manoharan[11], Ward Ortmann[11], Tushar Bhangale[11], Joshua C. Denny[12,13], Robert J. Carroll[12], Anne E. Eyler[13], Jeffrey D. Greenberg[14], Joel M. Kremer[15], Dimitrios A. Pappas[16], Lei Jiang[17], Jian Yin[17], Lingying Ye[17], Ding-Feng Su[18], Jian Yang[19,20], Gang Xie[21,22,23], Ed Keystone[24], Harm-Jan Westra[25], Tõnu Esko[3,26,27], Andres Metspalu[26], Xuezhong Zhou[28], Namrata Gupta[3], Daniel Mirel[3], Eli A. Stahl[29], Dorothée Diogo[1,2,3], Jing Cui[1,2,3], Katherine Liao[1,2,3], Michael H. Guo[1,3,27], Keiko Myouzen[10], Takahisa Kawaguchi[7], Marieke J. H. Coenen[30], Piet L. C. M. van Riel[31], Mart A. F. J. van de Laar[32], Henk-Jan Guchelaar[33], Tom W. J. Huizinga[34], Philippe Dieudé[35,36], Xavier Mariette[37], S. Louis Bridges Jr[38], Alexandra Zhernakova[25,34], Rene E. M. Toes[34], Paul P. Tak[39,40,41], Corinne Miceli-Richard[37], So-Young Bang[42], Hye-Soon Lee[42], Javier Martin[43], Miguel A. Gonzalez-Gay[44], Luis Rodriguez-Rodriguez[45], Solbritt Rantapää-Dahlqvist[46,47], Lisbeth Ärlestig[46,47], Hyon K. Choi[48,49,50], Yoichiro Kamatani[51], Pilar Galan[52], Mark Lathrop[53], the RACI consortium†, the GARNET consortium†, Steve Eyre[54,55], John Bowes[54,55], Anne Barton[54], Niek de Vries[56], Larry W. Moreland[57], Lindsey A. Criswell[58], Elizabeth W. Karlson[1], Atsuo Taniguchi[9], Ryo Yamada[59], Michiaki Kubo[60], Jun S. Liu[4], Sang-Cheol Bae[42], Jane Worthington[54,55], Leonid Padyukov[61], Lars Klareskog[61], Peter K. Gregersen[62], Soumya Raychaudhuri[1,2,3,63], Barbara E. Stranger[64,65], Philip L. De Jager[2,3,6], Lude Franke[25], Peter M. Visscher[19,20], Matthew A. Brown[19], Hisashi Yamanaka[9], Tsuneyo Mimori[8], Atsushi Takahashi[66], Huji Xu[17], Timothy W. Behrens[11], Katherine A. Siminovitch[21,22,23], Shigeki Momohara[9], Fumihiko Matsuda[7,67,68], Kazuhiko Yamamoto[10,69] & Robert M. Plenge[1,2,3]

[1]Division of Rheumatology, Immunology, and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. [2]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. [3]Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts 02142, USA. [4]Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, USA. [5]Centre for Cancer Research, Monash Institute of Medical Research, Monash University, Clayton, Victoria 3800, Australia. [6]Program in Translational NeuroPsychiatric Genomics, Institute for the Neurosciences, Department of Neurology, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. [7]Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto 606-8507, Japan. [8]Department of Rheumatology and Clinical immunology, Graduate School of Medicine, Kyoto University, Kyoto 606-8507, Japan. [9]Institute of Rheumatology, Tokyo Women's Medical University, Tokyo 162-0054, Japan. [10]Laboratory for Autoimmune Diseases, Center for Integrative Medical Sciences, RIKEN, Yokohama 230-0045, Japan. [11]Immunology Biomarkers Group, Genentech, South San Francisco, California 94080, USA. [12]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA. [13]Department of Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA. [14]New York University Hospital for Joint Diseases, New York, New York 10003, USA. [15]Department of Medicine, Albany Medical Center and The Center for Rheumatology, Albany, New York 12206, USA.

[16]Division of Rheumatology, Department of Medicine, New York, Presbyterian Hospital, College of Physicians and Surgeons, Columbia University, New York, New York 10032, USA. [17]Department of Rheumatology and Immunology, Shanghai Changzheng Hospital, Second Military Medical University, Shanghai 200003, China. [18]Department of Pharmacology, Second Military Medical University, Shanghai 200433, China. [19]University of Queensland Diamantina Institute, Translational Research Institute, Brisbane, Queensland 4072, Australia. [20]Queensland Brain Institute, The University of Queensland, Brisbane, Queensland 4072, Australia. [21]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario M5G 1X5, Canada. [22]Toronto General Research Institute, Toronto, Ontario M5G 2M9, Canada. [23]Department of Medicine, University of Toronto, Toronto, Ontario M5S 2J7, Canada. [24]Department of Medicine, Mount Sinai Hospital and University of Toronto, Toronto M5S 2J7, Canada. [25]Department of Genetics, University Medical Center Groningen, University of Groningen, Hanzeplein 1, Groningen 9700 RB, the Netherlands. [26]Estonian Genome Center, University of Tartu, Riia 23b, Tartu 51010, Estonia. [27]Division of Endocrinology, Children's Hospital, Boston, Massachusetts 02115, USA. [28]School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China. [29]The Department of Psychiatry at Mount Sinai School of Medicine, New York, New York 10029, USA. [30]Department of Human Genetics, Radboud University Medical Centre, Nijmegen 6500 HB, the Netherlands. [31]Department of Rheumatology, Radboud University Medical Centre, Nijmegen 6500 HB, the Netherlands. [32]Department of Rheumatology and Clinical Immunology, Arthritis Center Twente, University Twente & Medisch Spectrum Twente, Enschede 7500 AE, the Netherlands. [33]Department of Clinical Pharmacy and Toxicology, Leiden University Medical Center, Leiden 2300 RC, the Netherlands. [34]Department of Rheumatology, Leiden University Medical Center, Leiden 2300 RC, the Netherlands. [35]Service de Rhumatologie et INSERM U699 Hôpital Bichat Claude Bernard, Assistance Publique des Hôpitaux de Paris, Paris 75018, France. [36]Université Paris 7-Diderot, Paris 75013, France. [37]Institut National de la Santé et de la Recherche Médicale (INSERM) U1012, Université Paris-Sud, Rhumatologie, Hôpitaux Universitaires Paris-Sud, Assistance Publique-Hôpitaux de Paris (AP-HP), Le Kremlin Bicêtre 94275, France. [38]Division of Clinical Immunology and Rheumatology, Department of Medicine, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA. [39]AMC/University of Amsterdam, Amsterdam 1105 AZ, the Netherlands. [40]GlaxoSmithKline, Stevenage SG1 2NY, UK. [41]University of Cambridge, Cambridge CB2 1TN, UK. [42]Department of Rheumatology, Hanyang University Hospital for Rheumatic Diseases, Seoul 133-792, South Korea. [43]Instituto de Parasitologia y Biomedicina Lopez-Neyra, CSIC, Granada 18100, Spain. [44]Department of Rheumatology, Hospital Marques de Valdecilla, IFIMAV, Santander 39008, Spain. [45]Hospital Clinico San Carlos, Madrid 28040, Spain. [46]Department of Public Health and Clinical Medicine, Umeå University, Umeå SE-901 87, Sweden. [47]Department of Rheumatology, Umeå University, Umeå SE-901 87, Sweden. [48]Channing Laboratory, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston 02115, Massachusetts, USA. [49]Section of Rheumatology, Boston University School of Medicine, Boston, Massachusetts 02118, USA. [50]Clinical Epidemiology Research and Training Unit, Boston University School of Medicine, Boston, Massachusetts 02118, USA. [51]Centre d'Etude du Polymorphisme Humain (CEPH), Paris 75010, France. [52]Université Paris 13 Sorbonne Paris Cité, UREN (Nutritional Epidemiology Research Unit), Inserm (U557), Inra (U1125), Cnam, Bobigny 93017, France. [53]McGill University and Génome Québec Innovation Centre, Montréal, Québec H3A 0G1 Canada. [54]Arthritis Research UK Epidemiology Unit, Centre for Musculoskeletal Research, University of Manchester, Manchester Academic Health Science Centre, Manchester M13 9NT, UK. [55]National Institute for Health Research, Manchester Musculoskeletal Biomedical Research Unit, Central Manchester University Hospitals National Health Service Foundation Trust, Manchester Academic Health Sciences Centre, Manchester M13 9NT, UK. [56]Department of Clinical Immunology and Rheumatology & Department of Genome Analysis, Academic Medical Center/University of Amsterdam, Amsterdam 1105 AZ, the Netherlands. [57]Division of Rheumatology and Clinical Immunology, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, USA. [58]Rosalind Russell Medical Research Center for Arthritis, Division of Rheumatology, Department of Medicine, University of California San Francisco, San Francisco, California 94117, USA. [59]Unit of Statistical Genetics, Center for Genomic Medicine Graduate School of Medicine Kyoto University, Kyoto 606-8507, Japan. [60]Laboratory for Genotyping Development, Center for Integrative Medical Sciences, RIKEN, Yokohama 230-0045, Japan. [61]Rheumatology Unit, Department of Medicine (Solna), Karolinska Institutet, Stockholm SE-171 76, Sweden. [62]The Feinstein Institute for Medical Research, North Shore–Long Island Jewish Health System, Manhasset, New York 11030, USA. [63]NIHR Manchester Musculoskeletal Biomedical, Research Unit, Central Manchester NHS Foundation Trust, Manchester Academic Health Sciences Centre, Manchester M13 9NT, UK. [64]Section of Genetic Medicine, University of Chicago, Chicago, Illinois 60637, USA. [65]Institute for Genomics and Systems Biology, University of Chicago, Chicago, Illinois 60637, USA. [66]Laboratory for Statistical Analysis, Center for Integrative Medical Sciences, RIKEN, Yokohama 230-0045, Japan. [67]Core Research for Evolutional Science and Technology (CREST) program, Japan Science and Technology Agency, Kawaguchi, Saitama 332-0012, Japan. [68]Institut National de la Sante et de la Recherche Medicale (INSERM) Unite U852, Kyoto University Graduate School of Medicine, Kyoto 606-8507, Japan. [69]Department of Allergy and Rheumatology, Graduate School of Medicine, the University of Tokyo, Tokyo 113-0033, Japan. †Lists of participants and their affiliations appear in the Supplementary Information.

LETTER

doi:10.1038/nature12875

Selection and evaluation of clinically relevant AAV variants in a xenograft liver model

Leszek Lisowski[1]†, Allison P. Dane[2]†, Kirk Chu[1], Yue Zhang[1], Sharon C. Cunningham[2], Elizabeth M. Wilson[3], Sean Nygaard[4], Markus Grompe[4], Ian E. Alexander[2,5] & Mark A. Kay[1]

Recombinant adeno-associated viral (rAAV) vectors have shown early promise in clinical trials[1–3]. The therapeutic transgene cassette can be packaged in different AAV capsid pseudotypes, each having a unique transduction profile. At present, rAAV capsid serotype selection for a specific clinical trial is based on effectiveness in animal models. However, preclinical animal studies are not always predictive of human outcome[4–8]. Here, in an attempt to further our understanding of these discrepancies, we used a chimaeric human–murine liver model to compare directly the relative efficiency of rAAV transduction in human versus mouse hepatocytes *in vivo*. As predicted from preclinical and clinical studies[4,5,8], rAAV2 vectors functionally transduced mouse and human hepatocytes at equivalent but relatively low levels. However, rAAV8 vectors, which are very effective in many animal models, transduced human hepatocytes rather poorly—approximately 20 times less efficiently than mouse hepatocytes. In light of the limitations of the rAAV vectors currently used in clinical studies, we used the same murine chimaeric liver model to perform serial selection using a human-specific replication-competent viral library composed of DNA-shuffled AAV capsids. One chimaeric capsid composed of five different parental AAV capsids was found to transduce human primary hepatocytes at high efficiency *in vitro* and *in vivo*, and provided species-selected transduction in primary liver, cultured cells and a hepatocellular carcinoma xenograft model. This vector is an ideal clinical candidate and a reagent for gene modification of human xenotransplants in mouse models of human diseases. More importantly, our results suggest that humanized murine models may represent a more precise approach for both selecting and evaluating clinically relevant rAAV serotypes for gene therapeutic applications.

rAAV clinical trials have been hampered by unanticipated immunological responses and lower than expected levels of transgene product[5–8]. For a single serotype there is little correlation between *in vitro* and *in vivo* transduction of primary cells. Between two serotypes, AAV8 and AAV2, the former provides >10-fold higher levels of liver-mediated gene transfer in animals, including non-human primates[6,9–11]. An exception may be humans, where in the case of haemophilia B[6] the peak level of factor IX transgene product was similar in rAAV2- and rAAV8-treated individuals[8]. There are many potential reasons for the observed discordance in gene transfer efficiency among species, but relatively small differences in capsid sequence can affect both cellular uptake and post-cell-entry processing between species, ultimately affecting the level of gene transfer[12].

To establish if murine and human hepatocytes contained within the context of an intact liver are themselves differentially transduced, we injected $Fah^{-/-}/Rag2^{-/-}/Il2rg^{-/-}$ (FRG) mice[13] partially repopulated with primary human hepatocytes (hFRG) with single-stranded or self-complementary rAAV2 and rAAV8 vectors expressing enhanced green fluorescent protein (eGFP) (Fig. 1a, b). rAAV2 administration resulted in a low but equal number of eGFP-positive mouse and human hepatocytes. However, the rAAV8 vector resulted in a ~20-fold higher transduction efficiency in mouse compared with human hepatocytes, consistent with differences observed in preclinical and clinical studies published so far[6–8,10,11]. The block to functional rAAV8 transduction in human cells was not due to a lack of viral binding/uptake in the human hepatocytes because vector genomes were near equal in both human and mouse hepatocytes, as measured by quantitative polymerase chain reaction (qPCR) after laser capture microscopy (LCM) (Fig. 1c). By contrast, the rAAV2 vector genomes were preferentially taken up by human hepatocytes even though gene expression was similar in both the mouse and human cells. These results strongly suggest that differential functional transduction (measured as transgene expression) between capsid serotypes and species can be dependent on post-uptake factors.

Many different approaches have been used to alter the viral capsid and hence vector transduction properties[14–21]. As our goal was to identify new capsids with improved human tissue transduction, we created and screened a replicating AAV capsid library in the humanized mouse liver model. Our library screens are different from most in that selection is dependent not only on viral uptake and internalization, but also on viral replication, allowing us to select for these important post-uptake parameters that can affect vector-mediated gene transfer (reviewed in ref. 22).

Ten AAV capsid genes (AAV1, 2, 3B (ref. 23), 4, 5, 6, 8, 9, avian and bovine AAV) were used to generate an AAV-shuffled library (see Methods). To perform virus selection *in vivo* we used FRG mice partially repopulated with primary human hepatocytes (see Methods). Because the AAV libraries co-infected with wild-type human adenovirus 5 (hAd5) do not replicate in mice[14], we have a stringent simultaneous positive and negative selection between the human and murine cells, respectively.

We performed four rounds of selection (Fig. 2a) and monitored progress by sequencing >100 clones after each round (Fig. 2b). Library selection in non-humanized FRG animals in the presence of hAd5 served as a negative control and confirmed that rescued AAV was derived from the human cells (Fig. 2b), whereas a non-humanized FRG animal injected with hAd5 only served as a control to ensure that AAV *cap*-specific PCR signals were not caused by wild-type/rAAV contamination of the hAd5 preparation (Fig. 2a, b). After four rounds, the three most frequent variants, AAV-LK01, -LK02 and -LK03, accounted for 22.7%, 4.54% and 2.3% of the isolates, respectively. From the 19 most abundant variants, 15 were successfully used to package vectors, rAAV-RSV-eGFP (RSV, Rous sarcoma virus) (Fig. 2c). Reconstruction of the genealogical relationship at the DNA and amino acid level between the isolates and parental AAVs used to generate the library is shown in Fig. 2d.

[1]Stanford University, School of Medicine, Departments of Pediatrics and Genetics, 269 Campus Drive, Stanford, California 94305, USA. [2]Gene Therapy Research Unit, The Children's Hospital at Westmead and Children's Medical Research Institute, Locked Bag 4001, Westmead, 2145 New South Wales, Australia. [3]Yecuris Corporation, Portland, Oregon 97062, USA. [4]Oregon Stem Cell Center, Oregon Health and Science University, Portland, Oregon 97239, USA. [5]Discipline of Paediatrics and Child Health, The University of Sydney, 2145 New South Wales, Australia. †Present addresses: Gene Transfer, Targeting and Therapeutics Core, The Salk Institute for Biological Studies, 10010 N. Torrey Pines Rd, San Diego, California 92037, USA (L.L.); Department of Haematology, University College London Cancer Institute, London WC1E 6BT, UK (A.P.D.).

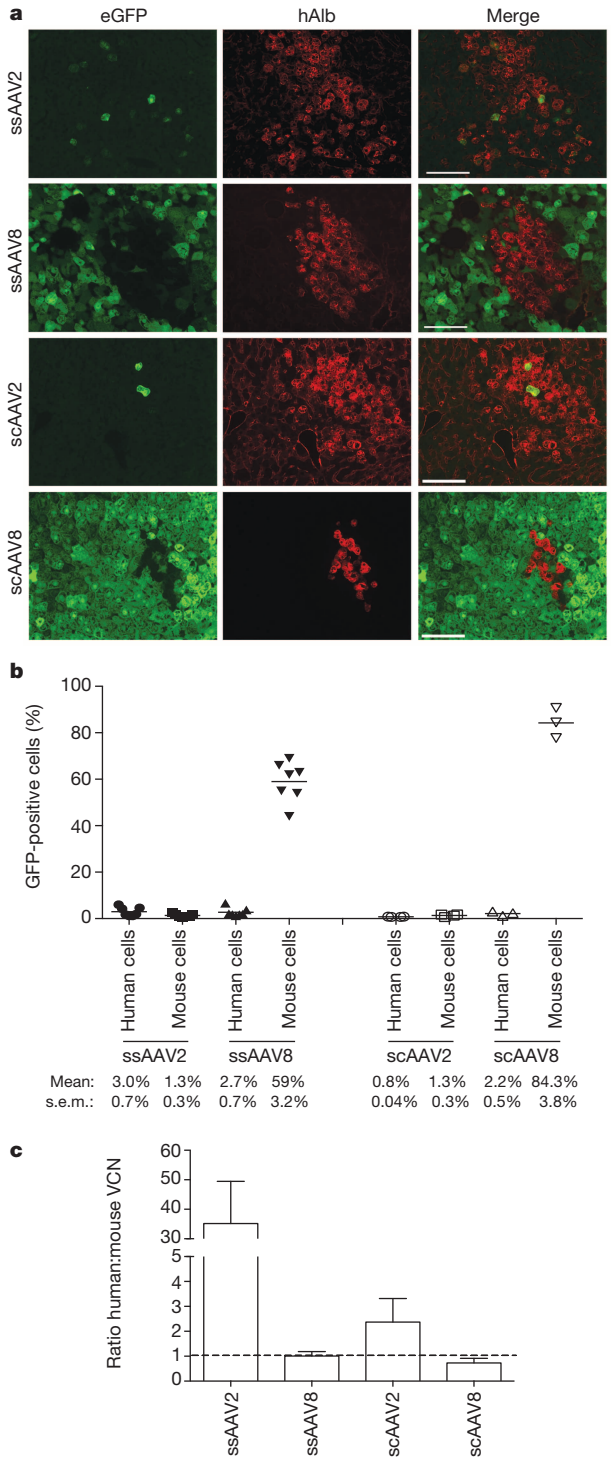


Figure 1 | *In vivo* comparison between rAAV2 and rAAV8. **a**, Representative histological images from the humanized FRG mouse livers transduced with 5×10^{10} vector genome (vg) single-stranded (ss) and self-complementary (sc) rAAV2 and rAAV8. The percentage of transduced human hepatocytes was determined by individually analysing and comparing cell counts from images of eGFP fluorescence and human albumin immunostaining (see Methods for details). hAlb, human albumin. Scale bar, 100 μm. **b**, Quantification of data shown in **a**. For ssAAV2 and ssAAV8, groups $n = 7$ each; for scAAV2, $n = 4$; for scAAV8, $n = 3$; up to 10 sections per mouse were scored. s.e.m., standard error of the mean. **c**, Ratio of vector genomes in human versus mouse hepatocytes *in vivo*. LCM followed by qPCR analysis was used to obtain relative vector copy numbers (VCNs) in each population (see Methods). Error bars show standard deviation (s.d.); $n = 3$.

We compared the *in vitro* transduction of these 15 AAV vector variants against standard AAV serotypes on a variety of cell types from different species (Extended Data Table 1). In comparison to rAAV-DJ, which efficiently transduces most mouse and human cells in culture[14], rAAV-LK03 preferentially transduced cells of human origin, suggesting that this variant has species-restricted specificity. Importantly, rAAV-LK03 transduced primary human hepatocytes 100-fold better than AAV8. AAV-LK03 is closely related to AAV3B, with lesser contributions from AAV1, 2, 4, 6, 8 and 9 (Figs 2d and 3a, Extended Data Table 2, Supplementary Table 1 and Supplementary Figs 1 and 2). rAAV-LK03 transduced human hepatocytes in culture 3, 67 and 6.5 times more efficiently than rAAV-DJ, rAAV8 and rAAV3B, respectively (Fig. 3b).

A second isolate, AAV-LK19, differed from AAV-LK03 by a single amino acid, S262C. This isolate transduced mouse cell lines and was less efficient than rAAV-LK03 at transducing most human cells, with the exception of primary human keratinocytes, where AAV-LK19 resulted in 5 and 7.6 times higher levels compared with rAAV-LK03 and rAAV6, the serotype reported to transduce human keratinocytes[24], respectively (Extended Data Table 1).

As pre-existing humoral immunity can block the clinical efficacy of a vector[8], we tested selected AAVs in a pooled human immunoglobulin-G (IVIG) neutralization assay (Extended Data Fig. 1 and Methods). Compared with AAV2, AAV3B, AAV-LK03 and AAV-LK19 were much more resistant to IVIG neutralization (Fig. 3c), suggesting that these vectors could transduce a large population of humans if used in a clinical trial[25,26].

Owing to the similarity between AAV-LK03, -LK19 and AAV3B, we wanted to establish whether the human hepatocyte growth factor (hHGF) receptor c-MET, previously identified as the AAV3 co-receptor[27], was involved in AAV-LK03 and -LK19 transduction. Transduction of Huh-7 cells with rAAV-hPGK-GFP-P2A-Luc2 (Luc2, luciferase 2) in the presence of increasing concentrations of hHGF (used for c-MET competition) demonstrated that rAAV-LK03, but not rAAV-LK19, was competitively inhibited by hHGF (Fig. 3d). These results suggest that the single amino acid difference (S262C) changes the receptor/co-receptor entry of AAV-LK19. Although the AAV capsid structure is available for several serotypes, the bulk of the sequence variation in these two new pseudotypes was in VP1 (amino acids 1–125), the one region of the capsid for which the structure has yet to be determined, making it difficult to speculate on specific structure–function relationships.

To compare the relative effectiveness of selected AAVs for transgene expression *in vivo*, we selected eight serotypes to package the hFIX expression cassette used in the rAAV2-FIX clinical trial[8]. In mice, the rAAV-LK03 and rAAV-LK19 vectors expressed very low levels of hFIX (Fig. 3e). By contrast, rAAV-LK01 and rAAV-LK02, which are closely related to AAV8 and AAV9, and AAV1 and AAV6, respectively (Fig. 2d), provided similar levels of expression as that observed with rAAV-DJ and rAAV8. Liver vector copy number (VCN) analysis performed 54 days after vector administration (Extended Data Fig. 2a) correlated with the hFIX expression data: rAAV8 $>$ rAAV-LK02 $\approx$ rAAV-DJ $\gg$ rAAV-LK03. Furthermore, liver AAV VCN determined at an early time point (Extended Data Fig. 2b) also correlated with transgene expression (Extended Data Fig. 2c), suggesting that rAAV-LK03 did not enter murine hepatocytes and/or was rapidly degraded upon cell entry.

To verify the strong species preference, we tested rAAV-LK03 in a hepatocellular carcinoma xenograft model[28]. Animals injected with rAAV3B showed no Luc expression in the liver or tumour, whereas animals injected with rAAV8 showed Luc expression from the tumour and the liver (Fig. 4a and Supplementary Figs 3, 4). Consistent with the data obtained with rAAV-hFIX in normal mice (Fig. 3e and Extended Data Fig. 2c), animals treated with rAAV-LK03-RSV-Luc2 did not show a detectable Luc signal from the liver but only from human tumours. In addition, the onset of transgene expression from AAV-LK03 was slower by about 48 h compared to AAV8 (Extended Data Fig. 3).

Having demonstrated that AAV-LK03 selectively transduced human cells, we next evaluated its transduction efficiency in FRG mice

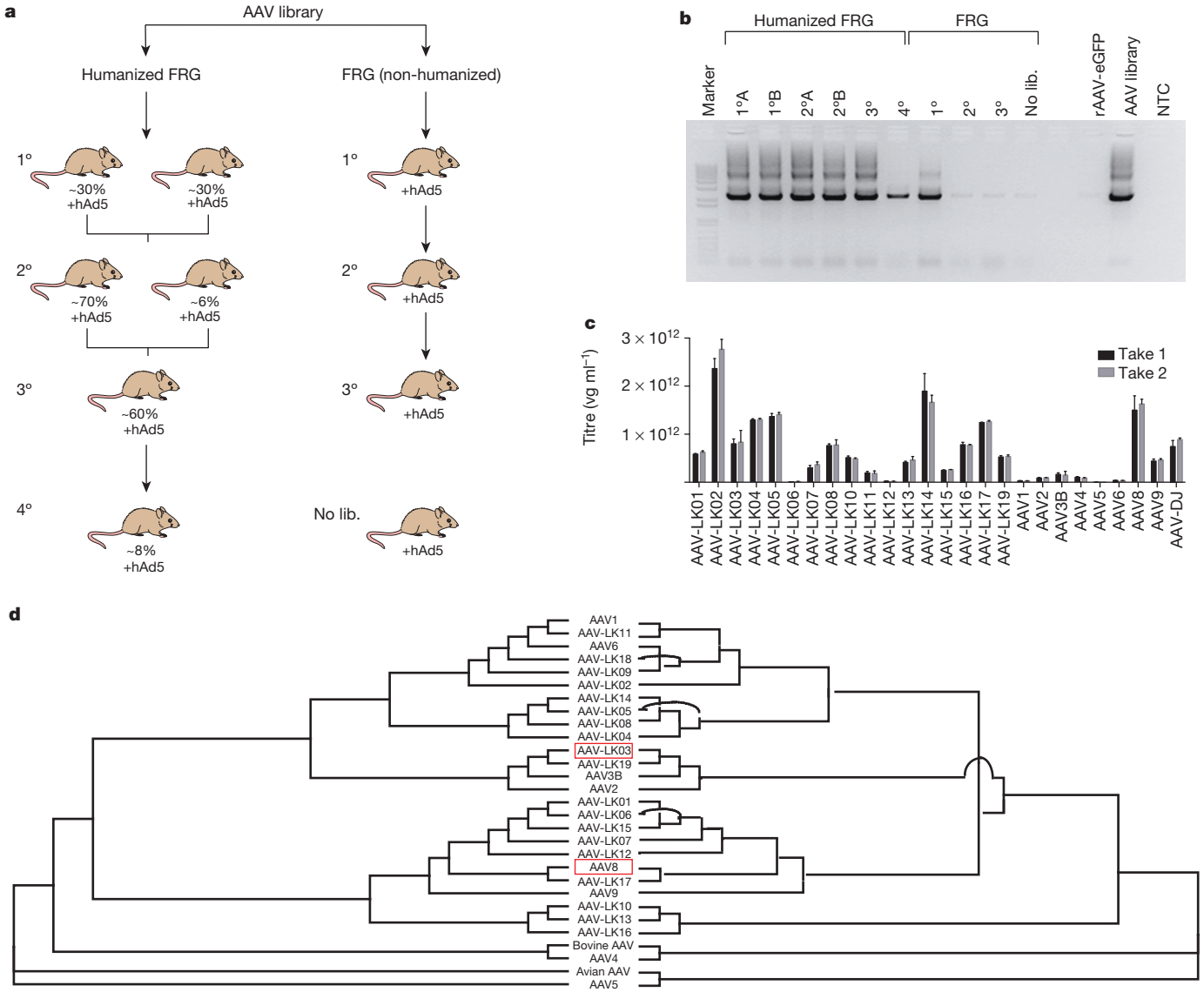


Figure 2 | *In vivo* AAV-shuffled library screen. **a**, Graphic representation of the screen. $N = 2$ for passages 1 (1°) and 2 (2°), $n = 1$ for passages 3 (3°) and 4 (4°). Percentage of human hepatocyte repopulation is listed. Selection was performed once (single biological repeat). No lib., no library control. **b**, PCR analysis of liver lysates after each selection step. NTC, no template control. Marker, '1 kb Plus' DNA ladder. A and B indicate mice in the duplicates (see Fig. 2a). **c**, Dot blot titre comparison between parental AAVs, AAV-DJ and novel isolates for two transduction experiments. Error bars show s.d.; $n = 3$. **d**, Genealogical relationship of isolates at protein (left) and DNA (right) levels.

with and without human hepatocyte reconstitution (Fig. 4b). In agreement with previously published data[27], rAAV3B did not transduce murine hepatocytes (control animals in Fig. 4b and Extended Data Fig. 4). Surprisingly, rAAV3B, previously shown to transduce human hepatocytes in culture[27] (Fig. 3b), did not lead to detectable Luc expression in engrafted human hepatocytes in this *in vivo* model. By contrast, a strong Luc signal was detected in non-humanized FRG animals injected with rAAV8-RSV-Luc (Fig. 4b and Extended Data Fig. 4), which efficiently transduced mouse hepatocytes *in vivo* (Fig. 3e and Extended Data Fig. 2). Most importantly, as predicted, the Luc signal was detected in humanized FRG mice injected with rAAV-LK03, but not in non-humanized FRG controls. These data showed that rAAV-LK03 selectively transduces human hepatocytes *in vivo*.

To quantify further the transduction efficiency of rAAV-LK03, we injected an equal dose of rAAV-LK03 or rAAV8-eGFP into humanized FRG mice. One week later, the number of transduced cells was compared (Fig. 4c). The fraction of GFP-positive human hepatocytes was 43.3 ± 11% and 3.6 ± 1.1% with rAAV-LK03 and AAV8 vector infusion, respectively. By contrast, almost all (>99%) mouse hepatocytes were eGFP positive with rAAV8, whereas ≪1% of mouse hepatocytes were eGFP positive after rAAV-LK03 vector infusion. Thus, rAAV-LK03 transduction, as measured by the number of transgene-positive cells, was about ten times higher than AAV8 in human hepatocytes *in vivo*. This experiment may underestimate the differences in AAV-mediated human hepatocyte transduction because of the variation in the rate of capsid uncoating and peak transgene expression among different serotypes. Whereas rAAV8 is rapidly uncoated[29], longer uncoating for rAAV-LK03 can be inferred from the slower rise in transgene expression in the tumour xenotransplant transduction studies (Extended Data Fig. 3).

The fact that there was not always a correlation between vector uptake and transgene expression between species (Fig. 1) strongly suggests that post-receptor entry factors (such as intracellular transport, nuclear entry and/or uncoating) influence the final level of functional transduction (reviewed in ref. 12). By making replication in human cells a condition for selection, we placed selection pressure on many of these post-entry parameters. Moreover, the human primary hepatocytes were in an environment that emulated their native

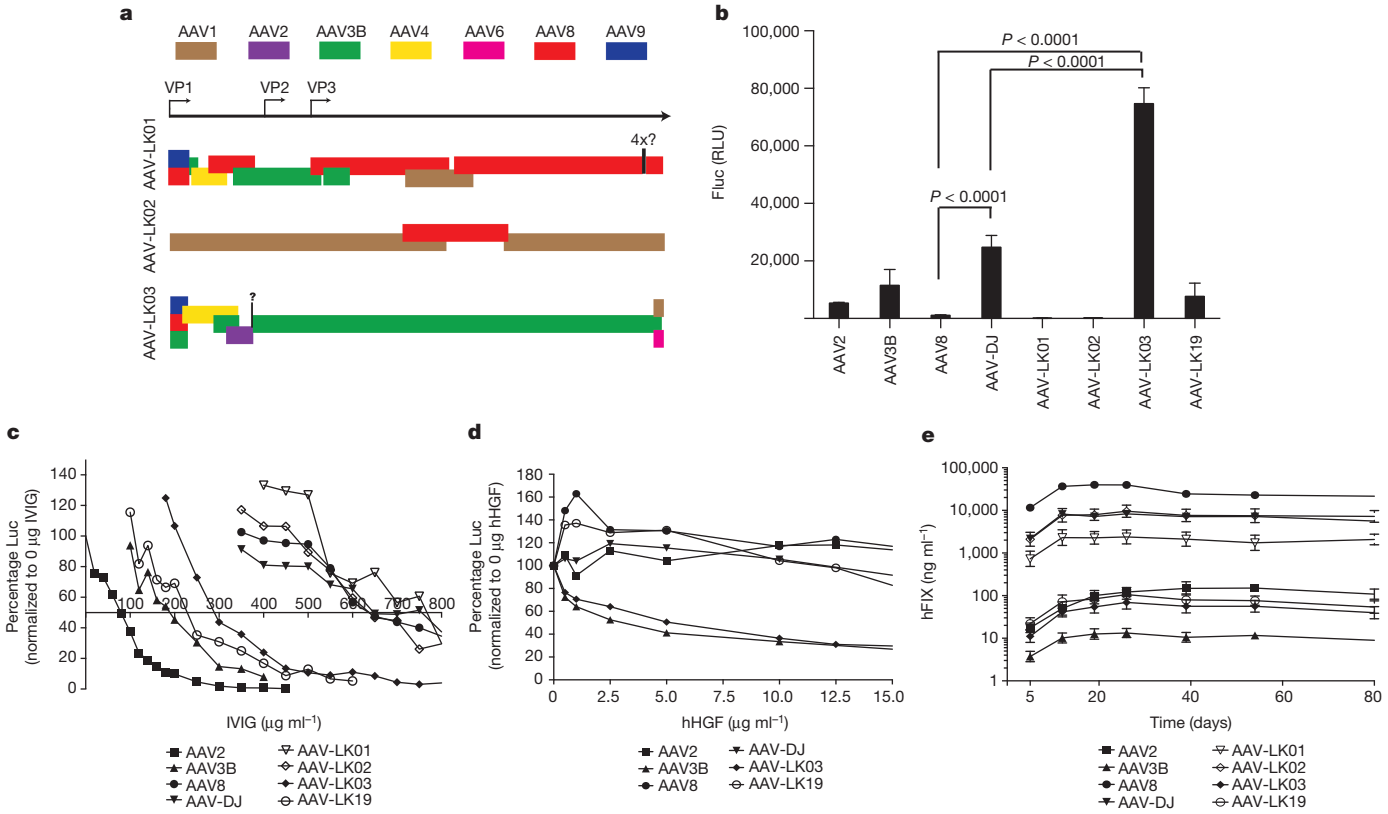


Figure 3 | Functional analysis of selected isolates. **a**, Contribution comparison for the top three isolates in graphic form. Residues with unknown parental origin are indicated with a question mark. **b**, Transgene expression in primary human hepatocytes in culture (see Methods). Fluc, firefly luciferase; RLU, relative light unit. Error bars represent s.d.; $n = 4$ from a single biological repeat. **c**, IVIG neutralization assay; $n = 4$ (<15% variation). **d**, hHGF competition assay. Data represent averages from two biological repeats with $n = 3$ per repeat. **e**, Comparison of hFIX expression in C57BL6 mice. $N = 5$ animals per group up to day 54, $n = 2$ at day 80. Single biological repeat (see also Extended Data Fig. 2c).

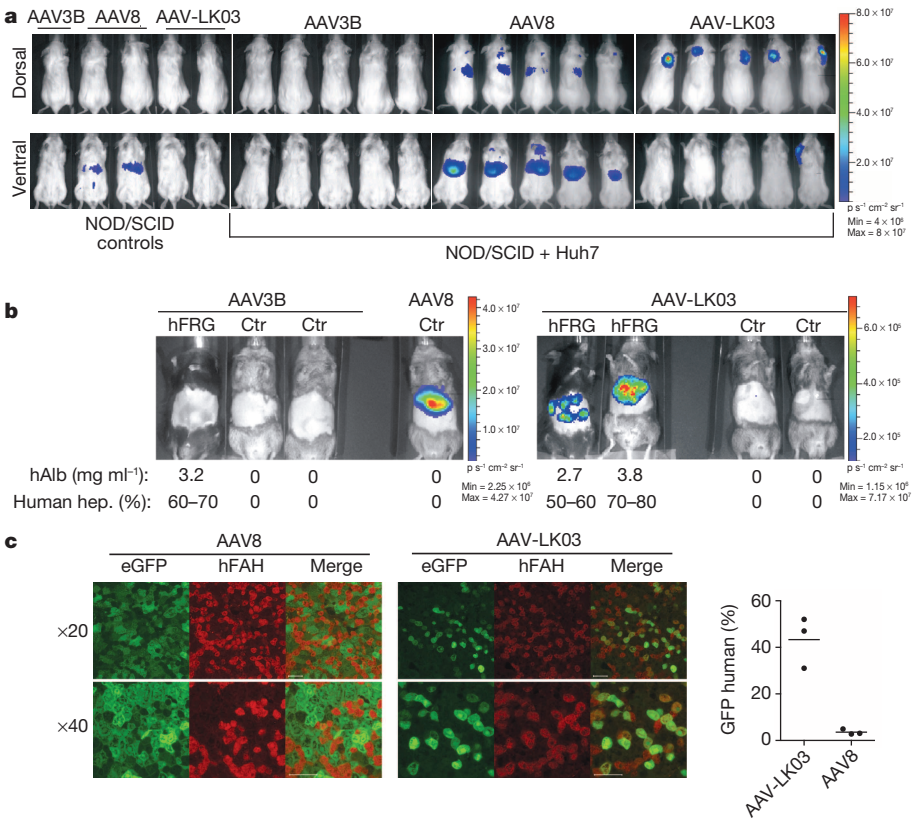


Figure 4 | *In vivo* vector specificity analysis. **a**, Luc expression in human hepatocellular carcinoma xenograft model 6 days after intravenous vector injection ($n = 5$ per group). Controls, naive animals. The same results were observed in two independent biological repeats. **b**, Luc expression in humanized FRG animals (hFRG) or naive controls (Ctr). Serum human albumin (hAlb) levels and estimated percentages of human hepatocyte (hep.) repopulation are given. All animals used in the study are shown: $n = 1+2$ (rAAV3B), $n = 1$ (rAAV8), $n = 2+2$ (rAAV-LK03). **c**, *In vivo* comparisons between hFRG animals transduced with 5×10^{11} vg rAAV8 or rAAV-LK03. Representative histological images are shown. Cell counting was as in Fig. 1a (see Methods). hFAH, human fumarylacetoacetate hydrolase. Scale bar, 100 μm. The graph represents quantification of *in vivo* data from three animals.

setting and thus maintained a more physiological gene expression profile compared with an *in vitro* setting. This made selecting a candidate useful in humans more likely. A summary of the transduction efficiencies of standard and newly selected AAV capsids described here is shown in Extended Data Table 3. Taken together, the rAAV-LK03-based vector is not only a highly promising clinical candidate, but will also prove useful when restricted genetic manipulation of xenotransplanted cells/tissues is required.

Most importantly, our studies suggest that the use of a human primary cell xenotransplant model compared with commonly used mouse and non-human primate models may more accurately predict potential transduction efficiency in humans. The approach provided herein may accelerate the ability to identify and establish clinically useful AAV vector candidates for clinical trials in which current serotype selection is in large part based on gene transfer efficiency in animal models.

METHODS SUMMARY

An AAV capsid library derived from ten different capsids was generated as previously described[14], with minor modifications. The library was subsequently injected into FRG mice reconstituted with primary human hepatocytes[13]. Twenty-four hours later, wild-type hAd5 was injected to induce replication of the AAV library in the human hepatocytes. Forty-eight hours after hAd5 injection, livers were harvested and analysed for capsid sequences by PCR and Sanger sequencing. These liver lysates were used for subsequent rounds of selection. Computer algorithms were designed for sequence comparisons and capsid of origin contribution analysis.

After four passages of the library, selected capsid clones were cloned into standard AAV vector production plasmids and then used to package various vector expression cassettes. rAAV-RSV-eGFP, RSV-Luc and hPGK-GFP-P2A-Luc expression cassettes packaged in various capsids were used in cell culture studies[14]. AAV-RSV-Luc and rAAV-LSP1-eGFP[30] were used in tumour xenotransplant and FRG *in vivo* studies, respectively. In the FRG mice, 1–2 weeks after vector infusion the number of human hepatocytes was determined by anti-human-albumin staining, and the transduced mouse and human cells were quantified by eGFP fluorescence[30]. In some animals, the number of proviral vector genomes in mouse versus human hepatocytes was determined by LCM and qPCR[30].

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 July; accepted 11 November 2013.
Published online 25 December 2013; corrected online 10 January 2014 (see full-text HTML version for details).

1. Bainbridge, J. W. *et al.* Effect of gene therapy on visual function in Leber's congenital amaurosis. *N. Engl. J. Med.* **358,** 2231–2239 (2008).
2. Gaudet, D. *et al.* Efficacy and long-term safety of alipogene tiparvovec (AAV1-LPL^{S447X}) gene therapy for lipoprotein lipase deficiency: an open-label trial. *Gene Ther.* **20,** 361–369 (2013).
3. Bennett, J. *et al.* AAV2 gene therapy readministration in three adults with congenital blindness. *Sci. Trans. Med.* **4,** 120ra115 (2012).
4. Nietupski, J. B. *et al.* Systemic administration of AAV8-α-galactosidase A induces humoral tolerance in nonhuman primates despite low hepatic expression. *Mol. Ther.* **19,** 1999–2011 (2011).
5. Jiang, H. *et al.* Effects of transient immunosuppression on adenoassociated, virus-mediated, liver-directed gene transfer in rhesus macaques and implications for human gene therapy. *Blood* **108,** 3321–3328 (2006).
6. Nathwani, A. C. *et al.* Adenovirus-associated virus vector-mediated gene transfer in hemophilia B. *N. Engl. J. Med.* **365,** 2357–2365 (2011).
7. Kay, M. A. *et al.* Evidence for gene transfer and expression of factor IX in haemophilia B patients treated with an AAV vector. *Nature Genet.* **24,** 257–261 (2000).
8. Manno, C. S. *et al.* Successful transduction of liver in hemophilia by AAV-Factor IX and limitations imposed by the host immune response. *Nature Med.* **12,** 342–347 (2006).
9. Nathwani, A. C. *et al.* Safe and efficient transduction of the liver after peripheral vein infusion of self-complementary AAV vector results in stable therapeutic expression of human FIX in nonhuman primates. *Blood* **109,** 1414–1421 (2007).
10. Nathwani, A. C. *et al.* Self-complementary adeno-associated virus vectors containing a novel liver-specific human factor IX expression cassette enable highly efficient transduction of murine and nonhuman primate liver. *Blood* **107,** 2653–2661 (2006).
11. Davidoff, A. M. *et al.* Comparison of the ability of adeno-associated viral vectors pseudotyped with serotype 2, 5, and 8 capsid proteins to mediate efficient transduction of the liver in murine and nonhuman primate models. *Mol. Ther.* **11,** 875–888 (2005).
12. Nonnenmacher, M. & Weber, T. Intracellular transport of recombinant adeno-associated virus vectors. *Gene Ther.* **19,** 649–658 (2012).
13. Azuma, H. *et al.* Robust expansion of human hepatocytes in *Fah*$^{-/-}$/*Rag2*$^{-/-}$/*Il2rg*$^{-/-}$ mice. *Nature Biotechnol.* **25,** 903–910 (2007).
14. Grimm, D. *et al. In vitro* and *in vivo* gene therapy vector evolution via multispecies interbreeding and retargeting of adeno-associated viruses. *J. Virol.* **82,** 5887–5911 (2008).
15. Müller, O. J. *et al.* Random peptide libraries displayed on adeno-associated virus to select for targeted gene therapy vectors. *Nature Biotechnol.* **21,** 1040–1046 (2003).
16. Perabo, L. *et al. In vitro* selection of viral vectors with modified tropism: the adeno-associated virus display. *Mol. Ther.* **8,** 151–157 (2003).
17. Maheshri, N., Koerber, J. T., Kaspar, B. K. & Schaffer, D. V. Directed evolution of adeno-associated virus yields enhanced gene delivery vectors. *Nature Biotechnol.* **24,** 198–204 (2006).
18. Li, W. *et al.* Engineering and selection of shuffled AAV genomes: a new strategy for producing targeted biological nanoparticles. *Mol. Ther.* **16,** 1252–1260 (2008).
19. Pulicherla, N. *et al.* Engineering liver-detargeted AAV9 vectors for cardiac and musculoskeletal gene transfer. *Mol. Ther.* **19,** 1070–1078 (2011).
20. Asuri, P. *et al.* Directed evolution of adeno-associated virus for enhanced gene delivery and gene targeting in human pluripotent stem cells. *Mol. Ther.* **20,** 329–338 (2012).
21. Yang, L. *et al.* A myocardium tropic adeno-associated virus (AAV) evolved by DNA shuffling and *in vivo* selection. *Proc. Natl Acad. Sci. USA* **106,** 3946–3951 (2009).
22. Choi, V. W., McCarty, D. M. & Samulski, R. J. AAV hybrid serotypes: improved vectors for gene delivery. *Curr. Gene Ther.* **5,** 299–310 (2005).
23. Rutledge, E. A., Halbert, C. L. & Russell, D. W. Infectious clones and vectors derived from adeno-associated virus (AAV) serotypes other than AAV type 2. *J. Virol.* **72,** 309–319 (1998).
24. Petek, L. M., Fleckman, P. & Miller, D. G. Efficient *KRT14* targeting and functional characterization of transplanted human keratinocytes for the treatment of epidermolysis bullosa simplex. *Mol. Ther.* **18,** 1624–1632 (2010).
25. Calcedo, R., Vandenberghe, L. H., Gao, G., Lin, J. & Wilson, J. M. Worldwide epidemiology of neutralizing antibodies to adeno-associated viruses. *J. Infect. Dis.* **199,** 381–390 (2009).
26. Boutin, S. *et al.* Prevalence of serum IgG and neutralizing factors against adeno-associated virus (AAV) types 1, 2, 5, 6, 8, and 9 in the healthy population: implications for gene therapy using AAV vectors. *Hum. Gene Ther.* **21,** 704–712 (2010).
27. Ling, C. *et al.* Human hepatocyte growth factor receptor is a cellular coreceptor for adeno-associated virus serotype 3. *Hum. Gene Ther.* **21,** 1741–1747 (2010).
28. Hazari, S. *et al.* Hepatocellular carcinoma xenograft supports HCV replication: a mouse model for evaluating antivirals. *World J. Gastroenterol.* **17,** 300–312 (2011).
29. Thomas, C. E., Storm, T. A., Huang, Z. & Kay, M. A. Rapid uncoating of vector genomes is the key to efficient liver transduction with pseudotyped adeno-associated virus vectors. *J. Virol.* **78,** 3110–3122 (2004).
30. Cunningham, S. C., Dane, A. P., Spinoulas, A., Logan, G. J. & Alexander, I. E. Gene delivery to the juvenile mouse liver using AAV2/8 vectors. *Mol. Ther.* **16,** 1081–1088 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by National Institutes of Health grants HL092096 and HL064274 to M.A.K. and DK048252 to M.G.; L.L. was supported in part by the Berry Fellowship Foundation; I.E.A. by Australian National Health and Medical Research Council (NHMRC) grant 1008021.

Author Contributions L.L. helped with study design, performed experiments and data analysis, prepared figures and the manuscript. A.D. performed some of the experiments and data analysis, and assisted in figure preparation and manuscript editing. K.C. helped in performing some of the experiments. Y.Z. performed some of the vector sequence analysis. S.C.C. performed some of the animal studies and assisted in manuscript editing. E.M.W. generated the human transplanted FRG mice in Fig. 4c. S.N. injected the animals and prepared tissues for the experiment in Fig. 4c. M.G. helped with establishing the FRG colony and provided advice on *in vivo* human hepatocyte repopulation. I.E.A. helped with study design and manuscript editing. M.A.K. helped with study coordination, manuscript writing and editing. All authors reviewed and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.A.K. (markay@stanford.edu).

LETTER

doi:10.1038/nature12891

Convergent evolution of a fused sexual cycle promotes the haploid lifestyle

Racquel Kim Sherwood[1]†*, Christine M. Scaduto[1]*, Sandra E. Torres[1]† & Richard J. Bennett[1]

Sexual reproduction is restricted to eukaryotic species and involves the fusion of haploid gametes to form a diploid cell that subsequently undergoes meiosis to generate recombinant haploid forms. This process has been extensively studied in the unicellular yeast *Saccharomyces cerevisiae*, which exhibits separate regulatory control over mating and meiosis. Here we address the mechanism of sexual reproduction in the related hemiascomycete species *Candida lusitaniae*. We demonstrate that, in contrast to *S. cerevisiae*, *C. lusitaniae* exhibits a highly integrated sexual program in which the programs regulating mating and meiosis have fused. Profiling of the *C. lusitaniae* sexual cycle revealed that gene expression patterns during mating and meiosis were overlapping, indicative of co-regulation. This was particularly evident for genes involved in pheromone MAPK signalling, which were highly induced throughout the sexual cycle of *C. lusitaniae*. Furthermore, genetic analysis showed that the orthologue of *IME2*, a 'diploid-specific' factor in *S. cerevisiae*[1,2], and *STE12*, the master regulator of *S. cerevisiae* mating[3,4], were each required for progression through both mating and meiosis in *C. lusitaniae*. Together, our results establish that sexual reproduction has undergone significant rewiring between *S. cerevisiae* and *C. lusitaniae*, and that a concerted sexual cycle operates in *C. lusitaniae* that is more reminiscent of the distantly related ascomycete, *Schizosaccharomyces pombe*. We discuss these results in light of the evolution of sexual reproduction in yeast, and propose that regulatory coupling of mating and meiosis has evolved multiple times as an adaptation to promote the haploid lifestyle.

Sexual reproduction is divided into two programs: gamete fusion, in which haploid cells merge to form a diploid cell, and gametogenesis, in which a specialized cell division, meiosis, generates recombinant haploid gametes. Sexual reproduction in *S. cerevisiae* involves mating between cells of opposite mating types, generating diploid products[5] that subsequently undergo meiosis upon nutritional limitation[6,7]. It is unclear, however, how representative this sexual cycle is of that in other yeast species. Here we examine regulation of sexual reproduction in the related hemiascomycete *Candida lusitaniae*, a member of the *Candida* clade of human pathogens[8]. *C. lusitaniae* was recently shown to have a complete sexual cycle despite lacking orthologues of key meiosis genes such as *IME1*, encoding the master transcriptional regulator of meiosis in *S. cerevisiae*[8–10]. In addition, whereas *S. cerevisiae* is predominantly diploid, *C. lusitaniae* cells preferentially exist in the haploid form and therefore exhibit a transient diploid state[8,9].

Transcriptional profiling of *C. lusitaniae* cells progressing through mating and meiotic programs was performed and compared to those of *S. cerevisiae*. In total, 406 genes were induced more than fourfold during *C. lusitaniae* mating, including highly conserved MAPK genes that regulate pheromone signalling in diverse fungal species[11–13] (Fig. 1a and Extended Data Fig. 1). Interestingly, we also observed elevated expression of several genes that are orthologues of 'meiosis-specific' genes in *S. cerevisiae*, including *SPO11*, *REC8* and *IME2* (refs 1, 14–16) (Fig. 1a). We similarly performed expression profiling on *C. lusitaniae* diploid **a**/α cells induced to enter meiosis. Cells became enlarged by 8 h, exhibited a morphological change by 12 h, and by 18–36 h dyad (two-cell) spores were formed, as is typical of this species[9] (Extended Data Fig. 2a–c). Profiling revealed that a total of 618 genes were induced during *C. lusitaniae* meiosis, compared to 480 genes during *S. cerevisiae* meiosis[17]. In *S. cerevisiae*, meiotic gene expression includes early, middle and late stages of expression[17]. We were similarly able to distinguish three temporal classes of meiotic gene expression in *C. lusitaniae* (Fig. 1b, d and Extended Data Fig. 2d). In total, we observed increased expression (>threefold) of 255 early genes, 307 middle genes and 56 late genes during *C. lusitaniae* meiosis (Fig. 1e). Thus, despite lacking an orthologue of *IME1*, many downstream components of meiosis are regulated in a similar stage-specific manner in *C. lusitaniae* as in *S. cerevisiae*.

The most striking aspect of *C. lusitaniae* meiosis was the expression of many genes whose orthologues are expressed specifically during mating in *S. cerevisiae*. In particular, multiple components of the pheromone MAPK signalling cascade were highly induced, including the terminal transcription factor *STE12* (Fig. 1c, f). In addition to MAPK genes, pheromone, pheromone receptor and pheromone processing genes were also induced in *C. lusitaniae* meiosis (Extended Data Fig. 1). We therefore surmised that MAPK signalling might have a role in regulating *C. lusitaniae* meiosis, in addition to its conserved function in directing cell–cell communication and conjugation.

Given that program-specific expression of many *S. cerevisiae* mating and meiosis genes does not occur in *C. lusitaniae* orthologues, genetic experiments were performed to determine whether this transcriptional rewiring has functional consequences for the sexual cycle. First, the role of *C. lusitaniae* Ime2 in mating and meiosis was analysed. *S. cerevisiae* Ime2 (inducer of meiosis 2) is a conserved serine/threonine kinase that acts in tandem with the cyclin-dependent kinase Cdk1 to promote meiosis[1,2,18,19]. As predicted on the basis of *IME2* function in *S. cerevisiae*, loss of *IME2* (*CLUG_00015*) in *C. lusitaniae* blocked meiosis; wild-type diploid cells formed haploid progeny with ~40% efficiency after 3 days, whereas haploids were formed by <1% of *ime2Δ/ime2Δ* cells (Fig. 2a). In addition, whereas wild-type diploids generated dyad spores, *ime2Δ/ime2Δ* mutants failed to sporulate (Fig. 2b). These results establish a conserved role for *IME2* in regulating meiosis in hemiascomycete yeast, even in species for which *IME1* is absent. Profiling of *C. lusitaniae ime2Δ/ime2Δ* mutants revealed that most early meiosis genes were still induced in this background, whereas induction of many middle and late meiosis genes was lost (Extended Data Fig. 3). We also note that induction of *NDT80* (*CLUG_05634*) and genes encoding cell-cycle regulators was compromised in the *ime2Δ/ime2Δ* mutant (Extended Data Fig. 3b). *NDT80* is responsible for the induction of middle meiosis genes in *S. cerevisiae*, where its expression is dependent on Ime2 (ref. 20). The diminished expression of *NDT80*, together with the loss of expression of cell-cycle genes, is probably responsible for the inability of *C. lusitaniae ime2Δ/ime2Δ* cells to proceed through meiosis.

We next addressed the role of *IME2* in *C. lusitaniae* mating. In contrast to *S. cerevisiae* , in which *IME2* has no role in mating, *C. lusitaniae*

[1]Department of Microbiology and Immunology, Brown University, 171 Meeting Street, Providence, Rhode Island 02912, USA. †Present addresses: Department of Microbial Pathogenesis, Yale University, 295 Congress Avenue, New Haven, Connecticut 06536-0812, USA (R.K.S.); University of California San Francisco, Tetrad Graduate Program, San Francisco, California 94158-2330, USA (S.E.T.).
*These authors contributed equally to this work.

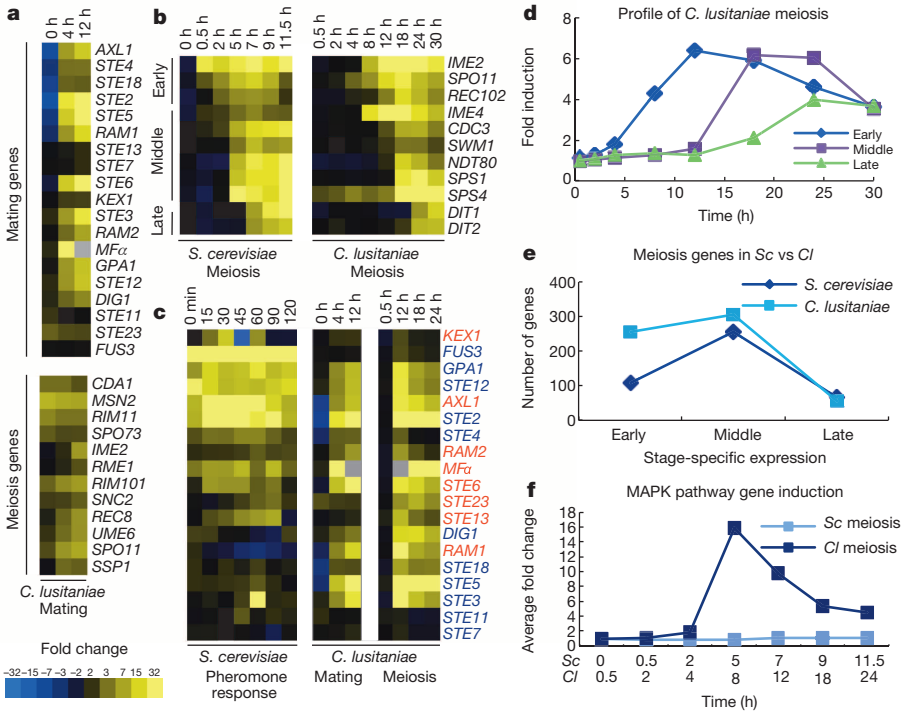


Figure 1 | Transcriptional profiling of mating and meiosis in *C. lusitaniae*. **a**, Profiling of *C. lusitaniae* mating. Top panel, induction of *C. lusitaniae* pheromone-signalling genes. Bottom panel, induction of genes characteristic of the meiotic program in *S. cerevisiae*. **b**, Stage-specific meiotic gene expression in *S. cerevisiae* and *C. lusitaniae*. Early, middle and late meiosis genes in *C. lusitaniae* are defined as those induced more than threefold after 12, 18 and 24 h, respectively. **c**, Expression of MAPK pathway genes (blue text) and pheromone-processing genes (red text) during *C. lusitaniae* mating and meiosis. Expression changes for *S. cerevisiae* genes[30] were enhanced threefold to be comparable to expression changes in *C. lusitaniae* genes. **d**, Average fold induction of early, middle and late meiosis genes in *C. lusitaniae*. **e**, Comparison of meiosis gene expression changes in *C. lusitaniae* (*Cl*) and *S. cerevisiae* (*Sc*). **f**, Fold induction of MAPK pathway genes during meiosis in *C. lusitaniae* and *S. cerevisiae*.

IME2 was induced during mating (Fig. 2c) and was required for the efficient formation of mating products. Thus, co-incubation of *C. lusitaniae* **a** and α cells resulted in approximately 0.5% of cells forming **a**/α diploids, whereas bilateral crosses between *ime2Δ* **a** and *ime2Δ* α cells resulted in a mating frequency of less than 0.03% (a 16-fold decrease, Fig. 2d). Unilateral crosses between wild type and *ime2Δ* mutants showed mating frequencies similar to those between wild-type partners (Fig. 2d). Complementation by reintegration of the *IME2* gene into the mutant background restored mating competency in the bilateral cross (Extended Data Fig. 4). These results establish a novel role for *IME2* in *C. lusitaniae* mating, and indicate that the *ime2Δ* defect is limited to bilateral crosses. The latter observation is consistent with a late role for Ime2 during mating following the fusion of **a** and α partner cells. Microscopic analysis of *ime2Δ* crosses showed the formation of zygote structures, further supporting a late role for Ime2 during mating (Fig. 2e), and additional studies will be required to address the precise role of Ime2 in this process.

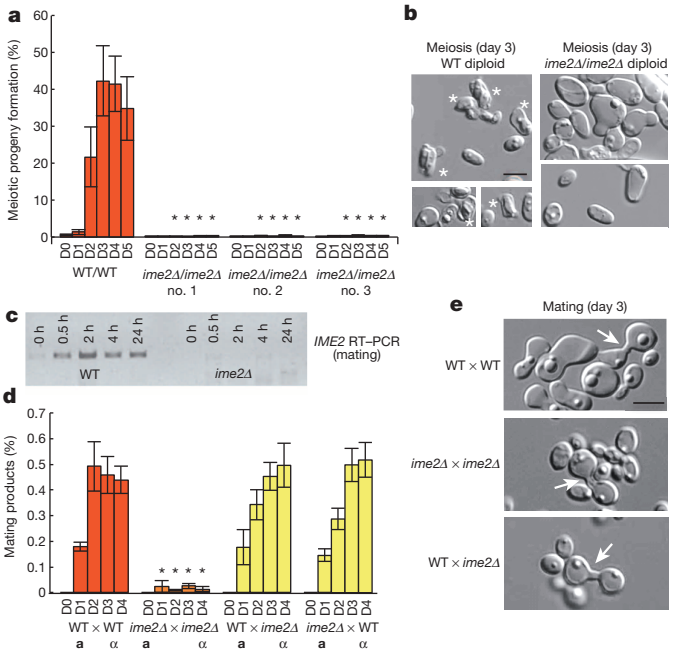


Figure 2 | *IME2* is required for both mating and meiosis in *C. lusitaniae*. **a**, Deletion of *IME2* blocks meiosis in *C. lusitaniae* (5-day time course, D1–D5). $*P < 0.01$, one way ANOVA, $n = 3$. WT, wild type. **b**, Wild-type diploid cells generate dyad spores (white asterisks) after 3 days on PDA medium, whereas *ime2Δ/ime2Δ* diploids do not sporulate, $n = 3$. **c**, PCR with reverse transcription (RT–PCR) data demonstrates induction of *IME2* during *C. lusitaniae* mating, $n = 3$. **d**, Loss of *IME2* in both **a** and α cells leads to decreased mating (4-day time course, D1–D4). $*P < 0.05$, Kruskal–Wallis, $n = 4$. **e**, *ime2Δ* mutants generate mating zygotes, indicating the block to mating occurs after cell fusion, $n = 4$. Scale bars, 5 μm. Data represented as mean ± s.e.m. All *n* values indicate number of biological replicates.

Next, we investigated the role of pheromone MAPK signalling and the transcription factor Ste12 in the *C. lusitaniae* sexual cycle. First, we tested whether *C. lusitaniae* cells undergoing meiosis actively secrete pheromone. A pheromone reporter assay was developed in which the *C. lusitaniae* α pheromone receptor was heterologously expressed in the related species *Candida albicans* (Fig. 3a). Such heterologous expression of receptors has been shown to result in successful signalling in response to species-specific pheromones[21]. Co-incubation of the yellow fluorescent protein (YFP)-labelled reporter strain with *C. lusitaniae* cells undergoing meiosis showed induction of a morphological response in the reporter cells (Fig. 3b), demonstrating that *C. lusitaniae* cells are actively secreting α pheromone during meiosis.

STE12 is the master regulator of mating and pheromone signalling in *S. cerevisiae* and many other hemiascomycetes, including *Candida* species[11,12]. To test the role of *STE12* in the sexual cycle of *C. lusitaniae* we constructed haploid and diploid *ste12Δ* mutants. Loss of *STE12* (*CLUG_02576*) in haploid cells led to a complete block in zygote formation and mating between **a** and α cells (Fig. 3c, d), as expected based on previous studies[22]. Surprisingly, however, loss of *STE12* in *C. lusitaniae* **a**/α diploids also led to a block in meiosis and sporulation. Thus, formation of meiotic progeny was reduced approximately 100-fold in *ste12Δ/ste12Δ* mutants (Fig. 3e) and spore formation was abolished (Fig. 3f and Extended Data Fig. 2c). Reintegration of the *STE12* gene into the *ste12Δ/ste12Δ* background restored the ability to undergo meiosis (Extended Data Fig. 5). Expression profiling revealed that the meiotic transcriptional

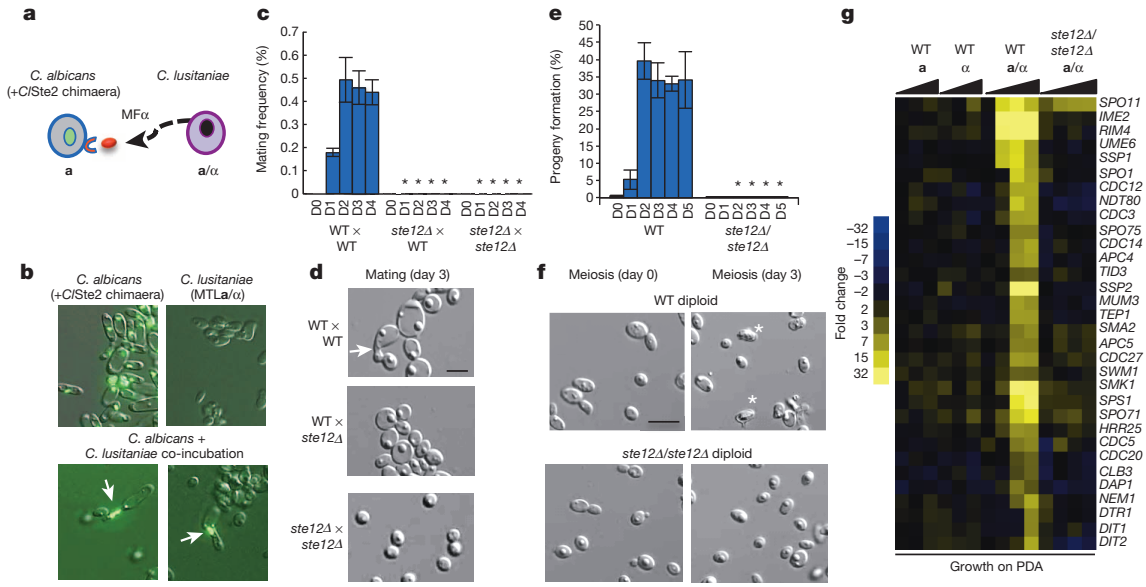


Figure 3 | *STE12* is essential for both mating and meiosis in *C. lusitaniae*. **a**, Schematic of assay to detect pheromone secretion from *C. lusitaniae* cells. *C. lusitaniae* **a**/α cells were co-incubated with YFP-labelled *C. albicans* **a** cells expressing the *C. lusitaniae* α pheromone receptor. **b**, *C. albicans* cells generate polarized mating projections (arrows) in response to *C. lusitaniae* pheromone, demonstrating that *C. lusitaniae* cells actively secrete pheromone during meiosis, $n = 2$. **c**, Deletion of *STE12* abolishes mating in *C. lusitaniae* (4-day time course, D1–D4). $^*P < 0.05$, Kruskal–Wallis, $n = 3$. **d**, Mating occurs between wild-type *C. lusitaniae* cells (arrow) but not between *ste12Δ* mutant cells, $n = 3$. **e**, Loss of *STE12* inhibited the formation of meiotic haploid progeny (4-day time course, D1–D4). $^*P < 0.05$, unequal variance *t*-test, $n = 3$. **f**, Absence of meiotic spores (white asterisks) in *C. lusitaniae* cells lacking *STE12*, $n = 5$. **g**, Gene expression changes when *C. lusitaniae* haploid, diploid or *ste12Δ/ste12Δ* strains are incubated on PDA medium. Deletion of *STE12* abolished most meiosis-specific gene expression changes. Where appropriate, scale bars, 5 μm; data represented as mean ± s.em. All *n* values indicate number of biological replicates.

profile was essentially absent from *ste12Δ/ste12Δ* mutants (Fig. 3g). Our results therefore establish that *STE12* has a critical role in regulating both mating and meiosis in *C. lusitaniae*, in contrast to *S. cerevisiae* where *STE12* specifically regulates mating only.

Taken together, our studies demonstrate that a fundamental rewiring of the sexual cycle has occurred between the two hemiascomycetes, *S. cerevisiae* and *C. lusitaniae*. Mating and meiosis are distinct programs in *S. cerevisiae*, yet these programs are integrated in *C. lusitaniae* (Fig. 4a and Extended Data Fig. 7). This is evident both from transcriptional profiles that reveal overlap of gene expression patterns in mating and meiosis, as well as genetic analysis of key regulators of the sexual cycle. In particular, whereas *STE12* and *IME2* are necessary for *S. cerevisiae* mating and meiosis, respectively, the *C. lusitaniae* orthologues are required for efficient progress through both stages of the sexual cycle.

The regulation of sexual reproduction in *C. lusitaniae* has notable parallels to that in the distantly related ascomycete *S. pombe*, even though these lineages diverged from one another more than 330 million years ago[23]. In *S. pombe*, mating and meiosis are also tightly coupled and both are dependent on pheromone MAPK signalling and the downstream transcriptional regulator, Ste11(refs 24, 25). In *C. lusitaniae*, we show that the pheromone MAPK-associated transcription factor *STE12* is similarly essential for both mating and meiosis. Thus, in both *S. pombe* and *C. lusitaniae* the programs regulating mating and meiosis are fused (Fig. 4a).

To address if Ste12 also regulates meiosis in other hemiascomycete species, we deleted the *STE12* orthologue from related sexual species including *S. cerevisiae*, *Kluyveromyces lactis*, *Pichia pastoris* and *Yarrowia lipolytica* (Extended Data Fig. 6). In each species loss of *STE12* failed to block meiosis, indicating that the meiotic function of *STE12* evolved relatively recently in the *C. lusitaniae* lineage (Fig. 4b). In support of this hypothesis we note that *C. lusitaniae*, but not the other hemiascomycete species, lost the transcriptional regulator α2 during evolution[8,9]. The *α2* gene acts to prevent expression of haploid-specific genes (including MAPK genes) in diploid **a**/α cells in diverse yeast species[26,27]. Thus, as

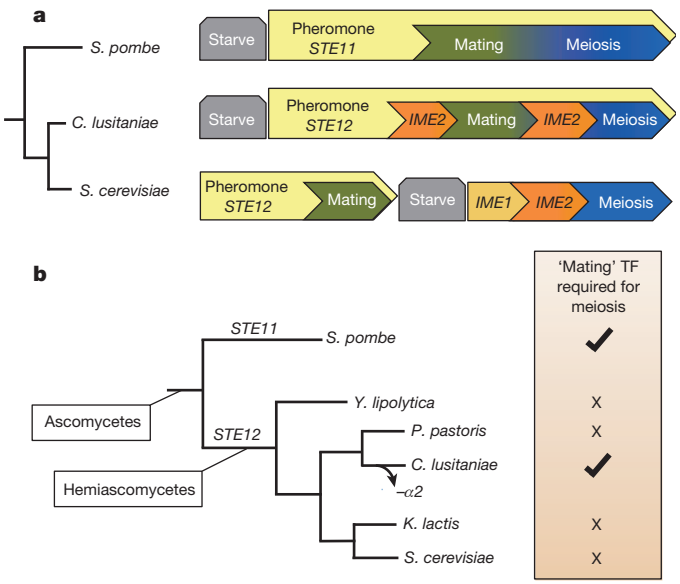


Figure 4 | Sexual regulation in yeast. **a**, Model comparing the sexual lifecycles of *S. cerevisiae*, *C. lusitaniae* and *S. pombe*. The diploid species *S. cerevisiae* shows distinct regulation of mating and meiosis; MAPK signalling via Ste12 regulates mating, whereas Ime1 regulates meiosis. In contrast, *S. pombe* shows integrated control of its sexual cycle; MAPK signalling and Ste11 regulate entry into both mating and meiosis. *C. lusitaniae* shows similar control over its sexual cycle to that in *S. pombe*. In particular, *C. lusitaniae* mating and meiosis are highly coupled, and both are regulated by the transcription factor Ste12. **b**, *C. lusitaniae* is the only hemiascomycete tested that requires *STE12* to undergo meiosis. Loss of the *MTLα2* transcription factor (TF) specifically in the *C. lusitaniae* lineage could have been a key step during evolution of this regulatory control, as it would permit expression of 'haploid-specific genes' (including MAPK genes) in diploid cells.

previously proposed[8,9], the loss of α2 could have preceded the rewiring of meiotic control in *C. lusitaniae*. In particular, we propose that it facilitated MAPK signalling and enabled Ste12 to assume control of meiosis in this species.

Why have such distinct modes of sexual regulation evolved in diverse unicellular yeast? The most parsimonious explanation is that these differences reflect a species' preference for one ploidy state over another. Both *C. lusitaniae* and *S. pombe*, despite being highly divergent, have haploid lifestyles that contrast sharply with the predominantly diploid lifestyle of *S. cerevisiae*. Co-regulation of the transcriptional programs controlling mating and meiosis therefore ensures that cells, once mated, immediately enter meiosis and return to the haploid state. This could represent an adaptive advantage for haploid cells in both *C. lusitaniae* and *S. pombe*. Alternatively, the selective pressure to keep these two sexual programs separate may have been lost during evolution, leading to the amalgamation of these programs in the two species. Although the relative advantages of haploidy and diploidy continue to be debated[28,29], it is clear that distinct transcriptional circuits can evolve to promote either haploid or diploid lifestyles. That this mode of regulation evolved independently in *S. pombe* and *C. lusitaniae* further suggests that examples of fused sexual cycles will be found throughout the fungal tree of life.

METHODS SUMMARY

Gene expression profiling. For mating and meiosis, *C. lusitaniae* cells were grown on YPD medium (non-inducing) or PDA medium (inductive for mating/meiosis). RNA was isolated from cells using the Ribopure-Yeast Kit (Life Technologies) and treated with Turbo DNase I (Ambion). Complementary DNA was synthesized with Oligo(dT)20, pdN9, and aa-dUTP/dNTPs using Superscript RT III. The resulting complementary DNAs were coupled to Cy3 and Cy5 and hybridized against a custom Agilent array for visualization.

Quantitative mating and meiosis assays. For mating assays, *C. lusitaniae* **a** and α cells were mixed on PDA medium for 1–5 days at 25 °C. Cells were subsequently analysed using medium selective for **a**, α, or **a**/α cells, and mating efficiencies calculated. For meiosis assays, diploid *C. lusitaniae* **a**/α cells were incubated on PDA medium for 1–5 days at 25 °C. Cells were plated on YPD medium and YPD+ cycloheximide to determine total number of cells and those that had undergone meiosis, respectively.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 July; accepted 18 November 2013.
Published online 5 January 2014.

1. Irniger, S. The Ime2 protein kinase family in fungi: more duties than just meiosis. *Mol. Microbiol.* **80,** 1–13 (2011).
2. Yoshida, M. *et al.* Initiation of meiosis and sporulation in *Saccharomyces cerevisiae* requires a novel protein kinase homologue. *Mol. Gen. Genet.* **221,** 176–186 (1990).
3. Dolan, J. W., Kirkman, C. & Fields, S. The yeast STE12 protein binds to the DNA sequence mediating pheromone induction. *Proc. Natl Acad. Sci. USA* **86,** 5703–5707 (1989).
4. Errede, B. & Ammerer, G. STE12, a protein involved in cell-type-specific transcription and signal transduction in yeast, is part of protein-DNA complexes. *Genes Dev.* **3,** 1349–1361 (1989).
5. Souza, C. A., Silva, C. C. & Ferreira, A. V. Sex in fungi: lessons of gene regulation. *Genet. Mol. Res.* **2,** 136–147 (2003).
6. van Werven, F. J. & Amon, A. Regulation of entry into gametogenesis. *Phil. Trans. R. Soc. Lond. B* **366,** 3521–3531 (2011).
7. Govin, J. & Berger, S. L. Genome reprogramming during sporulation. *Int. J. Dev. Biol.* **53,** 425–432 (2009).
8. Butler, G. *et al.* Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature* **459,** 657–662 (2009).
9. Reedy, J. L., Floyd, A. M. & Heitman, J. Mechanistic plasticity of sexual reproduction and meiosis in the *Candida* pathogenic species complex. *Curr. Biol.* **19,** 891–899 (2009).
10. Gargeya, I. B., Pruitt, W. R., Simmons, R. B., Meyer, S. A. & Ahearn, D. G. Occurrence of *Clavispora lusitaniae*, the teleomorph of *Candida lusitaniae*, among clinical isolates. *J. Clin. Microbiol.* **28,** 2224–2227 (1990).
11. Wong Sak Hoi, J. & Dumas, B. Ste12 and Ste12-like proteins, fungal transcription factors regulating development and pathogenicity. *Eukaryot. Cell* **9,** 480–485 (2010).
12. Lengeler, K. B. *et al.* Signal transduction cascades regulating fungal development and virulence. *Microbiol. Mol. Biol. Rev.* **64,** 746–785 (2000).
13. Butler, G. Fungal sex and pathogenesis. *Clin. Microbiol. Rev.* **23,** 140–159 (2010).
14. Keeney, S. in *Recombination and Meiosis* Vol. 2 (eds Egel, R. & Lankenau, D.) (Springer, 2007).
15. Klein, F. *et al.* A central role for cohesins in sister chromatid cohesion, formation of axial elements, and recombination during yeast meiosis. *Cell* **98,** 91–103 (1999).
16. Smith, H. E. & Mitchell, A. P. A transcriptional cascade governs entry into meiosis in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **9,** 2142–2152 (1989).
17. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282,** 699–705 (1998).
18. Guttmann-Raviv, N., Martin, S. & Kassir, Y. Ime2, a meiosis-specific kinase in yeast, is required for destabilization of its transcriptional activator, Ime1. *Mol. Cell. Biol.* **22,** 2047–2056 (2002).
19. Holt, L. J., Hutti, J. E., Cantley, L. C. & Morgan, D. O. Evolution of Ime2 phosphorylation sites on Cdk1 substrates provides a mechanism to limit the effects of the phosphatase Cdc14 in meiosis. *Mol. Cell* **25,** 689–702 (2007).
20. Sopko, R., Raithatha, S. & Stuart, D. Phosphorylation and maximal activity of *Saccharomyces cerevisiae* meiosis-specific transcription factor Ndt80 is dependent on Ime2. *Mol. Cell. Biol.* **22,** 7024–7040 (2002).
21. Lin, C. H., Choi, A. & Bennett, R. J. Defining pheromone-receptor signaling in *Candida albicans* and related asexual *Candida* species. *Mol. Biol. Cell* **22,** 4918–4930 (2011).
22. Young, L. Y., Lorenz, M. C. & Heitman, J. A *STE12* homolog is required for mating but dispensable for filamentation in *Candida lusitaniae*. *Genetics* **155,** 17–29 (2000).
23. Sipiczki, M. Where does fission yeast sit on the tree of life? *Genome Biol.* **1,** reviews1011 (2000).
24. Sugimoto, A., Iino, Y., Maeda, T., Watanabe, Y. & Yamamoto, M. *Schizosaccharomyces pombe ste11*$^{+}$ encodes a transcription factor with an HMG motif that is a critical regulator of sexual development. *Genes Dev.* **5,** 1990–1999 (1991).
25. Kjaerulff, S., Lautrup-Larsen, I., Truelsen, S., Pedersen, M. & Nielsen, O. Constitutive activation of the fission yeast pheromone-responsive pathway induces ectopic meiosis and reveals Ste11 as a mitogen-activated protein kinase target. *Mol. Cell. Biol.* **25,** 2045–2059 (2005).
26. Herskowitz, I. A regulatory hierarchy for cell specialization in yeast. *Nature* **342,** 749–757 (1989).
27. Booth, L. N., Tuch, B. B. & Johnson, A. D. Intercalation of a new tier of transcription regulation into an ancient circuit. *Nature* **468,** 959–963 (2010).
28. Otto, S. P. The evolutionary consequences of polyploidy. *Cell* **131,** 452–462 (2007).
29. Otto, S. P. & Gerstein, A. C. The evolution of haploidy and diploidy. *Curr. Biol.* **18,** R1121–R1124 (2008).
30. Roberts, C. J. *et al.* Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287,** 873–880 (2000).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Heitman, C. Lisset-Flores Mauriz, T. Noel, N. Hunter and J. Reedy for gifts of strains and plasmids, and S. Kabrawala and N. Balmuri for help with strain construction. We also thank T. Sorrells, L. Holt and members of the Johnson and Bennett laboratories for comments on the paper, and S. Jones for help with statistical analysis. This work was supported by National Science Foundation Grant MCB1021120 (to R.J.B.), by National Institutes of Health R01 Grant AI081704 (to R.J.B.), by T32GM007601 (to C.M.S.), by F31AI075607 (to R.K.S.), and by an Investigator in the Pathogenesis of Infectious Disease Award from the Burroughs Wellcome Fund (to R.J.B.).

Author Contributions R.K.S. and C.M.S. constructed strains, analysed phenotypes and performed transcriptional profiling experiments. S.E.T. and R.J.B. constructed strains and analysed phenotypes. R.K.S., C.M.S. and R.J.B. were involved in study design and writing of the manuscript.

Author Information All microarray data has been deposited into the NCBI Gene Expression Omnibus (GEO) portal under accession number GSE51794. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.J.B. (Richard_Bennett@brown.edu).

LETTER

doi:10.1038/nature12905

Structure of a *Naegleria* Tet-like dioxygenase in complex with 5-methylcytosine DNA

Hideharu Hashimoto[1], June E. Pais[2], Xing Zhang[1], Lana Saleh[2], Zheng-Qing Fu[3,4], Nan Dai[2], Ivan R. Corrêa Jr[2], Yu Zheng[2] & Xiaodong Cheng[1]

Cytosine residues in mammalian DNA occur in five forms: cytosine (C), 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC). The ten-eleven translocation (Tet) dioxygenases convert 5mC to 5hmC, 5fC and 5caC in three consecutive, Fe(II)- and α-ketoglutarate-dependent oxidation reactions[1–4]. The Tet family of dioxygenases is widely distributed across the tree of life[5], including in the heterolobosean amoeboflagellate *Naegleria gruberi*. The genome of *Naegleria*[6] encodes homologues of mammalian DNA methyltransferase and Tet proteins[7]. Here we study biochemically and structurally one of the *Naegleria* Tet-like proteins (NgTet1), which shares significant sequence conservation (approximately 14% identity or 39% similarity) with mammalian Tet1. Like mammalian Tet proteins, NgTet1 acts on 5mC and generates 5hmC, 5fC and 5caC. The crystal structure of NgTet1 in complex with DNA containing a 5mCpG site revealed that NgTet1 uses a base-flipping mechanism to access 5mC. The DNA is contacted from the minor groove and bent towards the major groove. The flipped 5mC is positioned in the active-site pocket with planar stacking contacts, Watson–Crick polar hydrogen bonds and van der Waals interactions specific for 5mC. The sequence conservation between NgTet1 and mammalian Tet1, including residues involved in structural integrity and functional significance, suggests structural conservation across phyla.

The free-living amoeboflagellate *Naegleria gruberi* has eight Tet/JBP-like dioxygenases (NgTet1–8; Extended Data Fig. 1). The NgTet proteins vary in length, but all contain a conserved core region of ~210 residues including the invariant Fe(II)-binding histidines and aspartate (the HxD…H motif). We measured NgTet1 activity using various double-stranded DNA as substrates, each containing a single modified base X within a G:X pair in a CpG sequence. We used antibodies specific for 5hmC, 5fC and 5caC (Extended Data Fig. 2a–c). Using 5mC-containing DNA as substrate, 5hmC (the first reaction product) and 5caC (the last reaction product) are detected in the presence of α-ketoglutarate (αKG), but not with *N*-oxalylglycine (NOG) (Fig. 1a). NgTet1 initially produces 5hmC at 5 min, 5fC between 5 to 10 min and finally 5caC at 15 min under the assay conditions (Fig. 1b). NgTet1 is active on all three DNA substrates containing 5mC, 5hmC or 5fC, generating 5caC (Fig. 1c). We applied quantitative mass spectrometry to monitor the kinetics of product formation (Fig. 1d and Extended Data Fig. 2d). When the amount of 5mC disappears rapidly (2–5 min), a peak of 5hmC forms transiently before being converted to 5fC and 5caC products (Fig. 1d). The first conversion from 5mC to 5hmC is faster ($k_{obs} = 21\ h^{-1}$) than the second conversion from 5hmC ($k_{obs} \approx 3\ h^{-1}$). In addition, we used human thymine DNA glycosylase to probe the products generated by NgTet1 (Extended Data Fig. 2e).

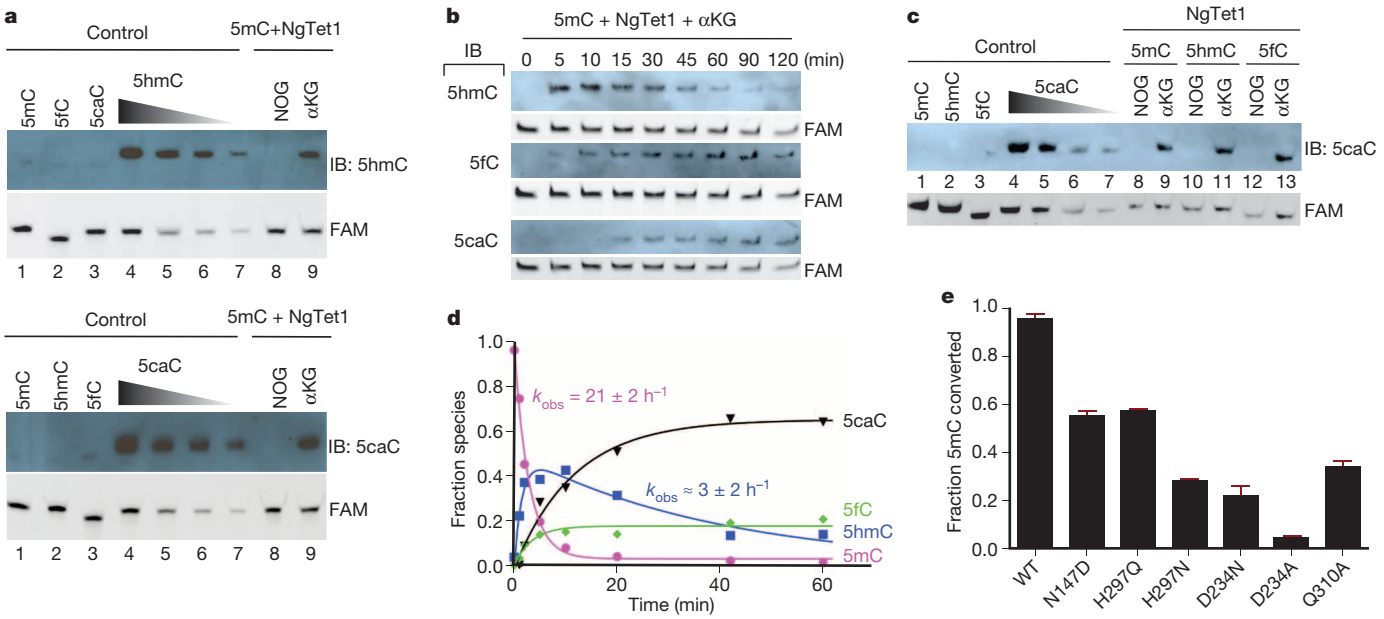


Figure 1 | Activity of NgTet1. **a**, Detection of 5hmC (top) and 5caC (bottom) by antibodies. FAM, fluorescein; IB, immunoblot. **b**, The relative amount of each reaction product was sequentially observed over the full time course of the reaction. **c**, NgTet1 is active on all three DNA substrates, producing 5caC. **d**, Quantitative LC–MS measurement of 5mC disappearance and formation of 5hmC, 5fC and 5caC. **e**, The effects of mutations on the conversion of 5mC. Error bars indicate s.e. of the mean value from three independent experiments. WT, wild type.

[1]Departments of Biochemistry, Emory University School of Medicine, 1510 Clifton Road, Atlanta, Georgia 30322, USA. [2]New England Biolabs, 240 County Road, Ipswich, Massachusetts 01938, USA. [3]Department of Biochemistry & Molecular Biology, University of Georgia, Athens, Georgia 30602, USA. [4]Sector 22, Advanced Photon Source, Argonne National Laboratory, Argonne, Illinois 60439, USA.

We determined the crystal structure of NgTet1 with a 14-base-pair (bp) oligonucleotide containing a single methylated CpG site in the presence of Mn^{2+} and NOG to form a catalytically inert complex, at 2.9 Å resolution (Extended Data Table 1). Like other structurally characterized αKG-dependent dioxygenases[8], NgTet1 has a core double-stranded β-helix fold that binds Fe(II) and αKG (Fig. 2a). Two twisted β-sheets (a four-stranded minor sheet and an eight-stranded major sheet) pack together with five helices on the outer surface of the major sheet to form a three-layered structure (Fig. 2a, b). The unequal number of strands of the two sheets creates an active site located asymmetrically on the side of the molecule where the extra strands of the major sheet are located. A 3_{10}-helix (h3 or h7) marks the end of each sheet and sits at the entrance to the active site. Two long loops associated with the 3_{10}-helices provide most of the functionally important residues. The hairpin loop (L1) between β5 and h3 of the major sheet recognizes the intrahelical guanine opposite to the target 5mC via Ser 148, and the extended loop (L2) connecting h7 from the minor sheet to the β7 of the major sheet (Fig. 2b) is responsible for binding of the metal ion (His 229 and Asp 231) and the flipped-out 5mC (Asp 234).

The DNA is bound to the basic surface of the protein with substantial protein-induced distortions from B-form DNA (Fig. 2c and Extended Data Fig. 3). The phosphate backbone flanking the CpG site is kinked ~65° and concurrently, one of the 5mC nucleotides flips out. Phosphate–protein contacts are concentrated on the four phosphates surrounding

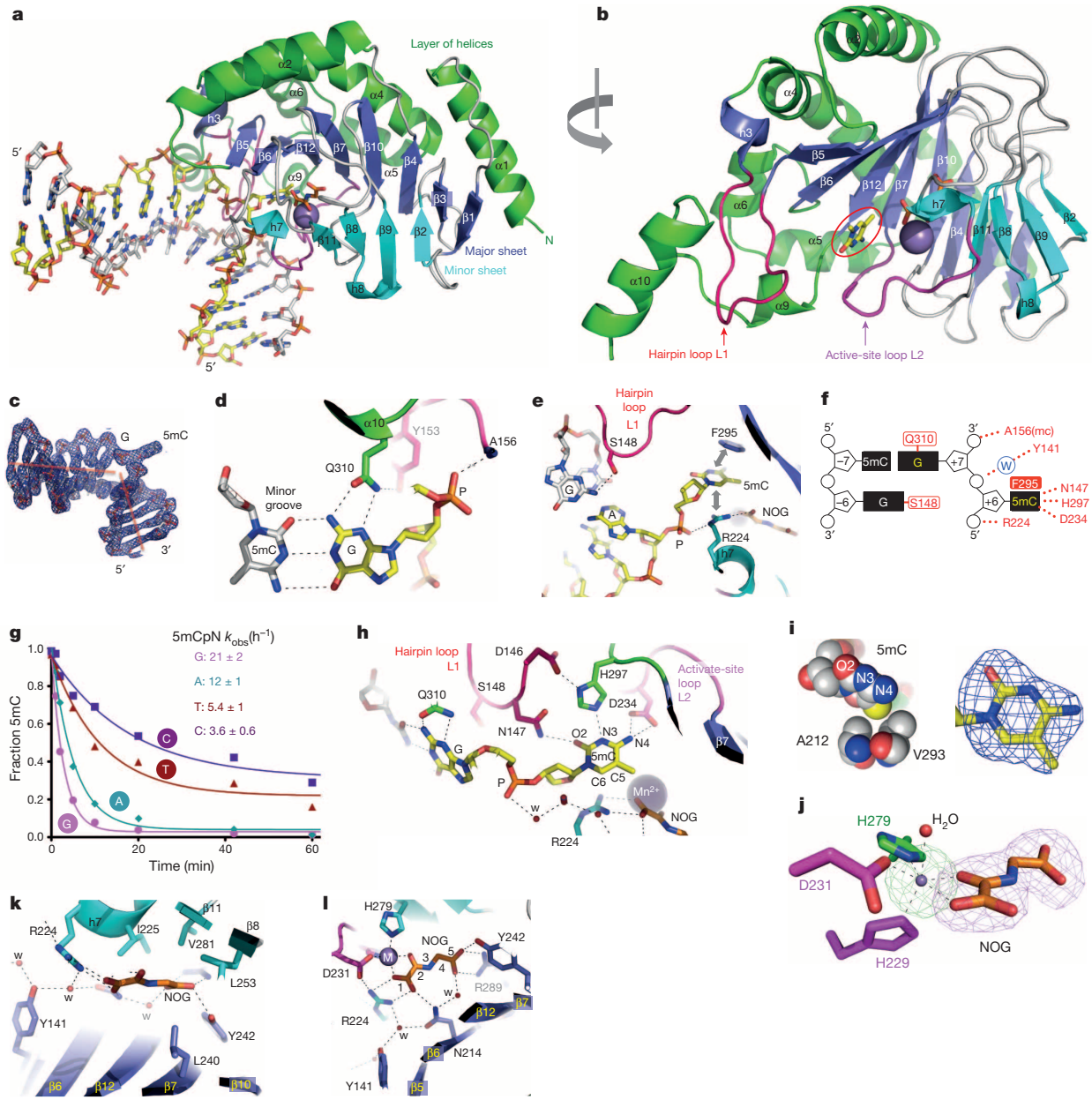


Figure 2 | Structure of NgTet1–DNA complex. a, The NgTet1 protein folds in a three-layered jelly-roll structure. **b**, Rotated ~90° from the view of panel **a**. **c**, Electron density $2F_o - F_c$, contoured at 1σ above the mean, is shown for the entire 14-bp DNA with a flipped out 5mC. **d**, Q310 interacts with 3′-G in the minor groove. **e**, S148 interacts with the intrahelical orphaned guanine. F295 and R224 form planar π stacking contacts with the extrahelical 5mC. **f**, Summary of the NgTet1–DNA interactions focusing on 5mCpG dinucleotide: mc, main-chain-atom-mediated contacts; W, water-mediated contacts. **g**, Substrate preference of 5mCpN (N = G, A, T or C) of NgTet1. **h**, The hydrogen-bond interactions with the polar atoms of 5mC. i, Left, the simulated annealing omit electron density, contoured at 4.5σ above the mean, for omitting 5mC and, right, the hydrophobic side chains of A212 and V293 border the methyl group (in yellow) of 5mC. Other atoms are coloured blue for nitrogen, red for oxygen and grey for carbon. **j**, The octahedral coordination of Mn^{2+} observed in the NgTet1–NOG–metal interactions (Extended Data Fig. 4f). Simulated annealing omit electron densities, contoured at 10σ and 5σ above the mean, are shown for Mn^{2+} (green mesh) and NOG (magenta mesh), respectively. **k, l**, Two views of NOG-NgTet1 interactions.

the flipped 5mC (Extended Data Fig. 3a, b), involving residues of the 3_{10}-helices h3 (Ala 156) and h7 (Arg 224) (Fig. 2d–f).

The enzyme approaches DNA from the minor groove, which is markedly widened near the flipped 5mC to ~10 Å in groove width owing to severe bending of the DNA. The tip of the hairpin loop, Ser 148, forms hydrogen bonds with the intrahelical orphaned guanine (Fig. 2e), whereas the side chain of Gln 310 of the carboxy-terminal helix α10 makes bifurcated hydrogen bonds with the 3′-guanine of the flipped 5mC (Fig. 2d). Such base-specific interactions would account for the preference of NgTet1 for 5mCpG as substrate. Replacing the 3′-guanine with adenine, thymine or cytosine resulted in reduction of the rate of 5mC conversion by a factor of ~1.75, 3.8 and 5.8, respectively (Fig. 2g). Similarly, mutating Gln 310 to alanine (Q310A) resulted in ~60% reduction of 5mC conversion (Fig. 1e). No direct interaction was observed for the 5mC in the opposite strand (Fig. 2d), consistent with NgTet1 being active on both fully and hemi-methylated CpG sites (Extended Data Fig. 2f).

The extrahelical 5mC is bound in a cage-like active site via stacking of the flipped base in between Phe 295 and the guanidino group of Arg 224 (Fig. 2e). Superimposition of a normal intrahelical 5mC onto the flipped 5mC indicates a very small rotation around the glycosidic bond (Extended Data Fig. 3d). The polar groups of the 5mC ring that normally form the Watson–Crick pairings with guanine now form hydrogen bonds with the side-chain amide group of Asn 147 (interacting with the O2 oxygen), the side-chain imidazole ring of His 297 (interacting with the N3 nitrogen), and the side-chain carboxylate oxygen atoms of Asp 234 (interacting with the N4 nitrogen) (Fig. 2h). Interactions with the exocyclic amino group N4 (NH_2) define the binding pocket specificity for a cytosine rather than thymine. Mutations of Asn 147, His 297 or Asp 234 resulted in much reduced (N147D, H297Q, H297N, D234N) or nearly abolished (D234A) activity on 5mCpG (Fig. 1e). The target methyl group—wedged between the hydrophobic side chains of Ala 212 and Val 293 (Fig. 2i)—is ~5.2 Å from the metal ion, which is similar to the observed distance (~4.5 Å) between the substrate atom to be oxidized and the iron in most structurally characterized αKG-oxygenases[8]. An additional hydroxyl, formyl or carboxylate group attached to the C5 methyl could fit into the space, consistent with 5hmC

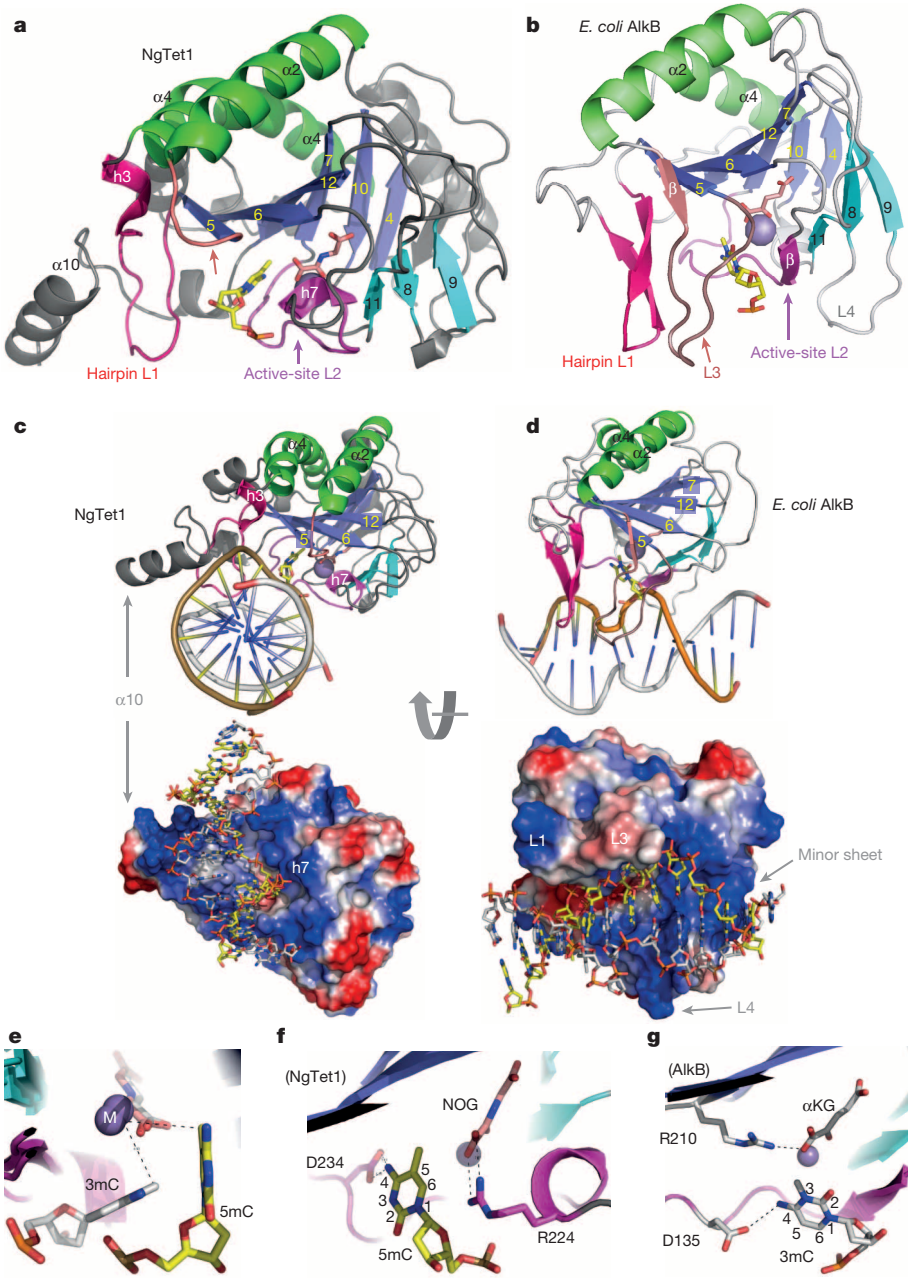


Figure 3 | Comparison of NgTet1 and AlkB. a, **b**, Structures of NgTet1 and AlkB aligned in a similar orientation. **c**, **d**, NgTet1 (**c**) and AlkB (**d**) are shown in relatively similar orientations. The surface charge at neutral pH is displayed as blue for positive, red for negative, and white for neutral. **e**, Superimposition of NgTet1 (5mC) and AlkB (3mC) in the active sites. The metal ions (M) are shown as balls and NOG or αKG (in the back) as sticks. **f**, **g**, Co-variation between the location of the target base (5mC in NgTet1 and 3mC in AlkB) and the NOG/αKG-interacting arginine (R224 of NgTet1 and R210 of AlkB).

or 5fC or 5caC being a substrate/product of NgTet1 (Extended Data Fig. 4a–e).

The metal ion Mn^{2+} has six ligands in an octahedral coordination (Fig. 2j). The NOG molecule is involved in extensive polar and hydrophobic interactions with the protein (Fig. 2k, l). The importance of these interactions is underscored by the fact that NOG-interacting residues are invariant or highly conserved among the eight NgTet-like homologues examined (Extended Data Fig. 1b). The NOG carboxylate group at the C5 position projects towards the interior hydrophobic core sandwiched between the two β-sheets (Fig. 2k), whereas the negatively charged carboxylate is balanced by the interaction with the invariant Arg 289 (Fig. 2l). The deep binding pocket of NOG (which is concealed in the NgTet1–DNA complex) indicates that the cofactor αKG binding precedes that of the DNA substrate (Extended Data Fig. 5), and stabilizes the NgTet1 structure by interacting with Arg 289 buried in the hydrophobic core.

The αKG dioxygenase family[8,9] includes members of the AlkB-like DNA/RNA repair enzymes[10]. We compared the complex structure of NgTet1–DNA–NOG–Mn^{2+} to that of *Escherichia coli* AlkB–DNA–αKG–Mn^{2+} (Fig. 3) and its human homologue ABH2 (Extended Data Fig. 6)[11,12] (the only other dioxygenases acting on nucleic acids structurally characterized in complex with DNA). The structures of NgTet1 and AlkB can be superimposed via the core elements of the jelly-roll fold (coloured in Fig. 3a, b). Both enzymes contain the hairpin loop (L1) after strand β5 and the active-site loop (L2) before strand β7. Besides the amino-terminal and C-terminal additions (Extended Data Fig. 6a), NgTet1 has, within the core region, extra helices α5 and α6, immediately after the kinked helix α4 (owing to Pro72 located in the middle of the helix). In the places of h3 and h7, two 3_{10}-helices unique to NgTet1 (Fig. 3a), AlkB has two additional β-strands, adjacent to β5 of the major sheet and β11 of the minor sheet, respectively (Fig. 3b). Unique to AlkB is an additional 12-residue-long loop (L3) before strand β5 making DNA backbone contacts, whereas the corresponding loop L3 in NgTet1 is a 4-residue short loop containing an invariant Lys 137 among the eight NgTet proteins (Extended Data Fig. 1c).

The most striking difference between NgTet1 and AlkB is that the bound DNA molecules lie nearly perpendicular to each other relative to the proteins (Fig. 3c, d). Both DNA molecules are bound against the basic surface of the protein (Fig. 3c, d), composed partly from the positively charged residues of the minor sheet unique to AlkB or the C-terminal helix α10 unique to NgTet1. We note that the C-terminal additions of all NgTet proteins (Extended Data Fig. 1b) and mammalian Tet enzymes are heavily enriched with basic residues that could also potentially interact with DNA. The vastly different protein–DNA interactions may reflect the fact that AlkB recognizes a damaged base pair, whereas NgTet1 recognizes a normal Watson–Crick base pair during the initial protein–DNA encounter. Like DNA methyltransferases[13] and DNA base excision repair enzymes[14], NgTet1 and AlkB (and ABH2) use a base-flipping mechanism to access the DNA bases where modification or repair occurs[15].

The perpendicular DNA-binding orientation also dictates how the flipped target base binds in the active site. The target nucleotide is simply rotated along the phosphodiester backbone (Extended Data Fig. 3d)[16], probably due to extensive protein–phosphate pinches[17] surrounding the flipped nucleotide. Thus, the flipped target bases, 5mC in NgTet1 and 3mC in AlkB, are also nearly perpendicularly positioned in their respective active sites (Fig. 3e). Yet, the distance between the target methyl group and the metal ion remains the same (~5 Å), consistent with a conserved chemical reaction. Also conserved is the ion-pair interaction of an active site arginine with the C1 carboxylate group of NOG of NgTet1 or αKG of AlkB, which is nearly superimposable (Extended Data Fig. 6c). However, the position of this arginine is different in the two enzymes, in accordance with the perpendicular orientation of the target bases (Fig. 3f, g). Therefore, the two enzymes approach the DNA substrates differently, resulting in distinct conformations of flipped target bases yet maintaining the ion-pair interaction with NOG/αKG.

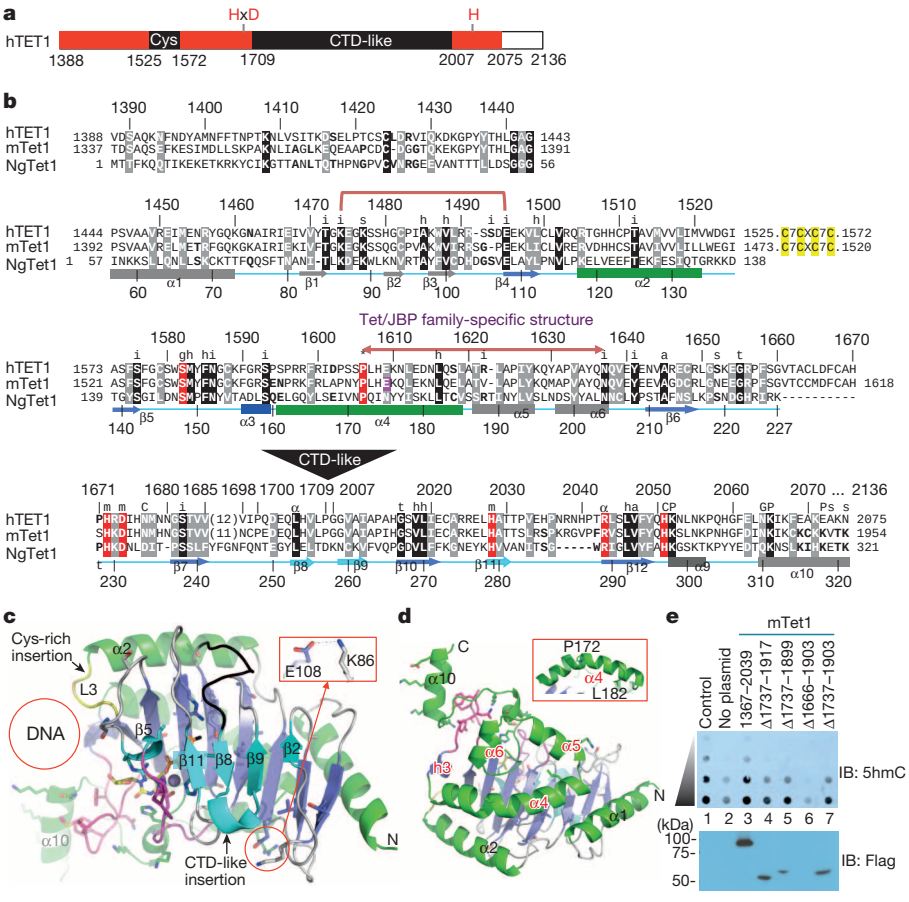


Figure 4 | Pairwise comparison of NgTet1 and mammalian Tet1. **a**, Schematic representation of hTet1 C-terminal catalytic domain. **b**, Sequence alignment of NgTet1, human Tet1 (hTet1) and mouse Tet1 (mTet1). Labels above the sequences indicate i for intra-molecular polar interaction; s for exposed surface residue; h for hydrophobic core; t for structural turn; α for αKG binding; m for metal ion coordination; P for DNA phosphate interaction; g for DNA base interaction with the orphaned guanine; G for DNA base interaction with the 3′ guanine to 5mC; C for 5mC interaction; a for active site residues (A212 and V293) near the methyl group of 5mC. **c**, Structure of NgTet1 with arrows indicating the two large insertions of mammalian Tet1. Highlighted is the charge–charge interaction between invariant K86 and E108. **d**, A kinked helix α4, owing to P172 (conserved among NgTet1, human and mouse Tet1, Tet2 and Tet3) located in the middle. **e**, Antibody detection of 5hmC in genomic DNA of HEK293T cells (top panel) expressing Flag-tagged mouse Tet1 catalytic domain or its internal deletions (bottom panel). Top panel, lane 1 is the 32-bp oligonucleotide containing a single 5hmC (20 pmol and twofold serial dilutions) and lanes 2–7 are the genomic DNA (500 ng and twofold serial dilutions). Bottom panel, lane 1 is the molecular weight marker and lanes 2 and 7 are the whole-cell lysates with approximately equal amount of protein.

Here we describe the first structure of a Tet-like dioxygenase, NgTet1, which is capable of converting 5mC to 5hmC, 5fC and 5caC. In mammalian genomes, the products of Tet enzymes include 5hmC in both CpG and non-CpG sequence context[18–20]. Likewise, NgTet1 is active on 5mCpG and 5mCpA (in a reduced rate). Structurally, NgTet1 represents the core structure of the catalytic domain of the mammalian Tet enzymes. The mammalian Tet proteins have their catalytic domains located in the C-terminal part of the proteins[1] with an atypical insertion of ~300 residues not found in other αKG-dioxygenases (Fig. 4a). The insertion separates the two halves of the ferrous binding motif, HxD…H. In addition, a stretch of ~50 residues containing a unique symmetrically spaced four cysteine residues CX_7CXCX_7C is located in the N-terminal portion of the catalytic domain. Removing these two insertions shows that NgTet1 and mammalian Tet1 share ~14% identity or ~39% similarity (Fig. 4b), the highest conservation among the pairwise comparisons of NgTet1 and other αKG-oxygenases examined (Extended Data Table 2 and Extended Data Figs 6 and 7). The sequence conservation is scattered throughout the entire region, including the residues involved in structural integrity and those with functional significance (DNA binding, base-specific interactions, metal ion and αKG bindings) (Extended Data Table 3). The conservation extends beyond the core of the jelly-roll fold shared by NgTet1 and AlkB/ABH2 (for example, Lys 86–Glu 108 ion pair in Fig. 4c), indicating that NgTet1 and mammalian Tet1 share an overall higher degree of structural conservation, owing to their common substrate and enzymatic properties. Another structural conservation between NgTet1 and mammalian Tet1 involves an invariant proline located in the middle of helix α4, causing a kink (Fig. 4d)—a unique feature which might be conserved among the Tet/JBP family as the kinked helix α4, together with the following helices α5 and α6, is composed of a stretch of residues predicted to be Tet/JBP specific[1]. No corresponding helices α5 and α6 are present in other structurally characterized αKG-oxygenases examined (Extended Data Figs 6 and 7).

The two large insertions of mammalian Tet1 lie in the loop (L3) between helix α2 and strand β5 (Cys-rich region) and the loop between strands β8 and β9 (Fig. 4c). The Cys-rich insertion is in the DNA-binding interface and thus might have roles in DNA binding. The large 300-residue insertion, which shares significant sequence similarity to the C-terminal domain (CTD) of RNA polymerase II[21], points away from the catalytic core and has a potential regulatory function. Like mammalian Tet proteins, histone lysine-specific demethylase LSD1 has an atypical insertion of the Tower domain into the catalytic amine oxidase domain, whereas the closely related LSD2 is devoid of the insertion and is active[22]. Similarly, deletions of the CTD-insertion in mouse Tet1 catalytic domain retain activity when expressed in HEK293T cells (Fig. 4e), providing further support of the evolutionary conservation between NgTet1 and mammalian Tet proteins.

METHODS SUMMARY

We designed synthetic NgTet1 by optimizing the codon set for *Escherichia coli* and assembling the gene by overlapping oligonucleotides. We generated a hexahistidine–SUMO (small ubiquitin-like modifier)-tagged construct containing full-length NgTet1 (pXC1010). The tag was cleaved and the NgTet1 was crystallized with fully methylated 14 bp DNA in the presence of $MnCl_2$ and NOG. The structure was determined by single anomalous diffraction using bromine-labelled DNA. The 5mC dioxygenase activity of NgTet1 was assayed by three methods including specific antibodies, base excision by thymine DNA glycosylase (TDG) and liquid chromatography–mass spectrometry.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 22 May; accepted 19 November 2013.
Published online 25 December 2013.

1. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324,** 930–935 (2009).
2. Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466,** 1129–1133 (2010).
3. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333,** 1300–1303 (2011).
4. He, Y. F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333,** 1303–1307 (2011).
5. Iyer, L. M., Zhang, D., Maxwell Burroughs, A. & Aravind, L. Computational identification of novel biochemical systems involved in oxidation, glycosylation and other complex modifications of bases in DNA. *Nucleic Acids Res.* **41,** 7635–7655 (2013).
6. Fritz-Laylin, L. K. *et al.* The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. *Cell* **140,** 631–642 (2010).
7. Iyer, L. M., Abhiman, S. & Aravind, L. Natural history of eukaryotic DNA methylation systems. *Prog. Mol. Biol. Transl. Sci.* **101,** 25–104 (2011).
8. Aik, W., McDonough, M. A., Thalhammer, A., Chowdhury, R. & Schofield, C. J. Role of the jelly-roll fold in substrate binding by 2-oxoglutarate oxygenases. *Curr. Opin. Struct. Biol.* **22,** 691–700 (2012).
9. McDonough, M. A., Loenarz, C., Chowdhury, R., Clifton, I. J. & Schofield, C. J. Structural studies on human 2-oxoglutarate dependent oxygenases. *Curr. Opin. Struct. Biol.* **20,** 659–672 (2010).
10. Trewick, S. C., Henshaw, T. F., Hausinger, R. P., Lindahl, T. & Sedgwick, B. Oxidative demethylation by *Escherichia coli* AlkB directly reverts DNA base damage. *Nature* **419,** 174–178 (2002).
11. Yang, C. G. *et al.* Crystal structures of DNA/RNA repair enzymes AlkB and ABH2 bound to dsDNA. *Nature* **452,** 961–965 (2008).
12. Yi, C. *et al.* Iron-catalysed oxidation intermediates captured in a DNA repair dioxygenase. *Nature* **468,** 330–333 (2010).
13. Klimasauskas, S., Kumar, S., Roberts, R. J. & Cheng, X. Hhal methyltransferase flips its target base out of the DNA helix. *Cell* **76,** 357–369 (1994).
14. Slupphaug, G. *et al.* A nucleotide-flipping mechanism from the structure of human uracil-DNA glycosylase bound to DNA. *Nature* **384,** 87–92 (1996).
15. Roberts, R. J. & Cheng, X. Base flipping. *Annu. Rev. Biochem.* **67,** 181–198 (1998).
16. Horton, J. R. *et al.* Caught in the act: visualization of an intermediate in the DNA base-flipping pathway induced by Hhal methyltransferase. *Nucleic Acids Res.* **32,** 3877–3886 (2004).
17. Werner, R. M. *et al.* Stressing-out DNA? The contribution of serine-phosphodiester interactions in catalysis by uracil DNA glycosylase. *Biochemistry* **39,** 12585–12594 (2000).
18. Sun, Z. *et al.* High-resolution enzymatic mapping of genomic 5-hydroxymethylcytosine in mouse embryonic stem cells. *Cell Rep.* **3,** 567–576 (2013).
19. Yu, M. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149,** 1368–1380 (2012).
20. Ficz, G. *et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473,** 398–402 (2011).
21. Upadhyay, A. K., Horton, J. R., Zhang, X. & Cheng, X. Coordinated methyl-lysine erasure: structural and functional linkage of a Jumonji demethylase domain and a reader domain. *Curr. Opin. Struct. Biol.* **21,** 750–760 (2011).
22. Fang, R. *et al.* LSD2/KDM1B and its cofactor NPAC/GLYR1 endow a structural and molecular model for regulation of H3K4 demethylation. *Mol. Cell* **49,** 558–570 (2013).

Acknowledgements We thank R. J. Roberts who initiated this collaborative work, and participated both in the work and the writing of the manuscript. We thank J. R. Horton for critical comments and B. Baker for synthesizing the oligonucleotides. Y.Z. thanks C. Fulton for helpful discussions on *N. gruberi* biology. The Department of Biochemistry of Emory University School of Medicine supported the use of SER-CAT beamlines. This work was supported by grants from the National Institutes of Health GM049245 to X.C. (who is a Georgia Research Alliance Eminent Scholar) and GM095209 and GM105132 to Y.Z.

Author Contributions H.H. performed antibody-based and TDG-based activity assays, crystallographic experiments and expression of mouse Tet1 in HEK293T cells. X.Z. made the overexpression construct in *E. coli*, developed (together with H.H.) assay conditions and performed NgTet1-8 sequence analysis. J.E.P. and L.S. performed kinetic assays using the LC–MS method and J.E.P. characterized the mutants. Z.-Q.F. performed crystallographic phasing calculations and generated an initial poly-alanine model. N.D. and I.R.C. developed the LC–MS method for detection of modified cytosine residues. X.Z., Y.Z. and X.C. organized and designed the scope of the study, and all were involved in analysing data and preparing the manuscript.

Author Information The X-ray structures (coordinates and structure factor files) of NgTet1 with bound DNA have been submitted to PDB under accession number 4LT5. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to X.C. (xcheng@emory.edu) and/or Y.Z. (zhengy@neb.com).

CORRIGENDUM

doi:10.1038/nature12937

Corrigendum: Probabilistic cost estimates for climate change mitigation

Joeri Rogelj, David L. McCollum, Andy Reisinger, Malte Meinshausen & Keywan Riahi

Nature **493,** 79–83 (2013); doi:10.1038/nature11787

In Fig. 2c and d of this Letter, the case of 'Delayed action until 2020' under an intermediate level of future energy demand (green dashed line) previously contained an inconsistency in its evolution of global greenhouse gas emissions until 2020 (that is, its greenhouse gas emissions until 2020 were in line with the low-energy instead of the intermediate-energy demand case). This resulted in lower cumulative emissions and, hence, relatively higher probabilities of staying below a particular global-mean temperature threshold than what would otherwise be consistent with the intermediate-energy demand case in Table 1. Figure 2 has been corrected in the HTML and PDF versions of the Letter online. In addition, data for the corrected case ('Delayed action until 2020') have been updated accordingly in Supplementary Figs 2, 3, 4, 5, 8 and 9, and the Supplementary Data. None of these changes affect our results or conclusions.

CORRIGENDUM

doi:10.1038/nature13004

Corrigendum: Obesity–induced gut microbial metabolite promotes liver cancer through senescence secretome

Shin Yoshimoto, Tze Mun Loo, Koji Atarashi, Hiroaki Kanda, Seidai Sato, Seiichi Oyadomari, Yoichiro Iwakura, Kenshiro Oshima, Hidetoshi Morita, Masahira Hattori, Kenya Honda, Yuichi Ishikawa, Eiji Hara & Naoko Ohtani

Nature **499,** 97–101 (2013); doi:10.1038/nature12347

In this Letter, the forename of author Masahira Hattori was spelled incorrectly as 'Masahisa'. It has been corrected in the HTML and PDF versions of the manuscript.

CORRIGENDUM

doi:10.1038/nature13009

Corrigendum: DWARF 53 acts as a repressor of strigolactone signalling in rice

Liang Jiang, Xue Liu, Guosheng Xiong, Huihui Liu, Fulu Chen, Lei Wang, Xiangbing Meng, Guifu Liu, Hong Yu, Yundong Yuan, Wei Yi, Lihua Zhao, Honglei Ma, Yuanzheng He, Zhongshan Wu, Karsten Melcher, Qian Qian, H. Eric Xu, Yonghong Wang & Jiayang Li

Nature **504,** 401–405 (2013); doi:10.1038/nature12870

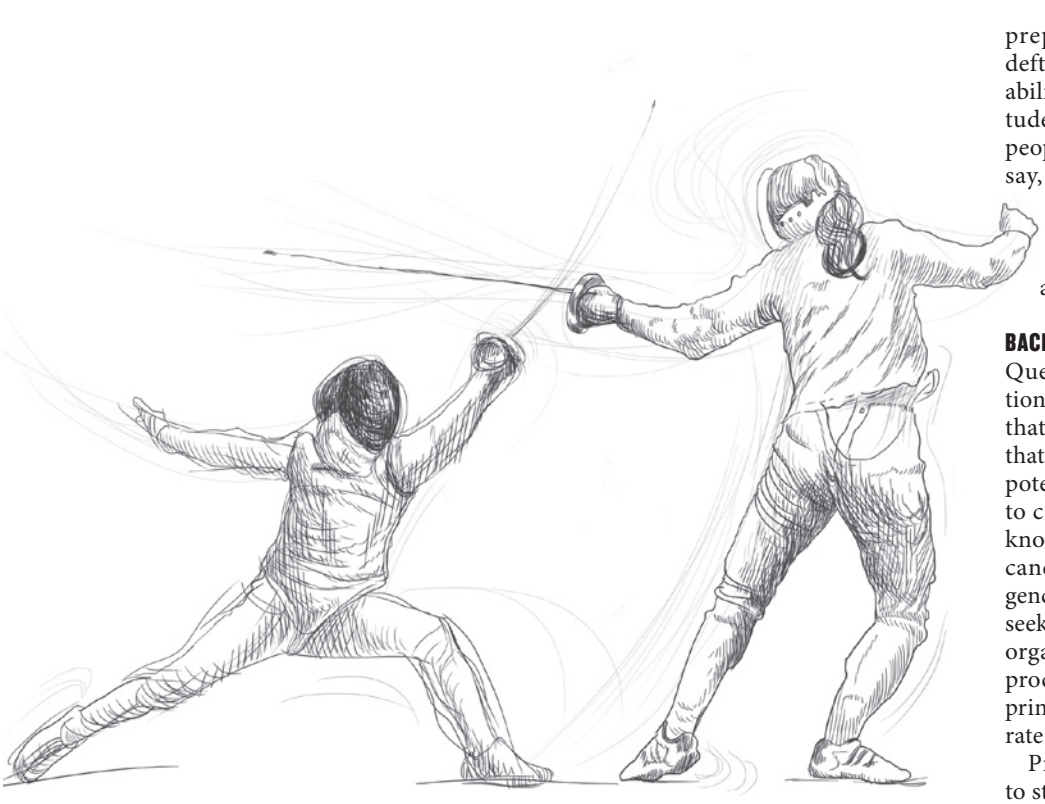
In this Article, the GenBank accession number was wrongly given as KF23088; it should be KF623088. The accession number has been corrected in the online versions of the paper.

CAREERS

MENTORING Protégés receive scant guidance on work–life conflicts **p.399**

DEGREES Enrolment slows in professional master's programme **p.399**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**



KUCO/SHUTTERSTOCK

INTERVIEWS

En garde

Navigating discussions with potential employers requires preparation and maintaining grace under pressure.

BY KAREN KAPLAN

It was the interview moment that everyone dreads — the killer question, then the pregnant pause. With a PhD in environmental biotechnology and years of experience as an environmental researcher and consultant, Henry Roman knew his subject well. Before his interview with the South African government's Department of Science and Technology he had spent hours reviewing that country's environmental legislation and international agreements on climate change.

But none of that stopped his mind from going blank when he was asked how he might develop a South African water-use policy. "I was coming in from a pure science background, and I'd had no experience in a policy environment," he says.

Yet Roman kept his cool. He asked his interviewer for half a minute to gather his thoughts, breathed and pulled it all together. "I drew on all the legislation I was aware of and on relevant international treaties to put the policy question into an international context," he recalls. He is now that department's director.

Invariably, early-career scientists who are on the interview circuit for any position will find themselves confronting a knotty interview question that they have no clue how to answer (see 'The 1-2-3 of interviews'). Hiring managers and other veteran interviewers say that, at such times, success depends on a hotchpotch of factors: ample advance preparation, excellent communication, deft interpersonal skills and a finely honed ability to keep calm. But above all, the attitude of the candidate is key. "I don't think people are censured for specific things they say, but because of the way they say them — nonchalant or arrogant or dispassionate," says Amy Cheng Vollmer, a microbiologist and department chair at Swarthmore College in Pennsylvania.

BACKGROUND CHECK

Questions that seek speculative information about what a candidate can accomplish, that ask about extracurricular activities or that broach seemingly extraneous topics are potentially flummoxing. But they are likely to cause less of a stumble when recipients know to expect them. Before an interview, candidates should gather as much intelligence as possible about the position they are seeking and about the institution, agency or organization that is hiring. This discovery process might uncover particulars about a principal investigator's priorities or a corporate drugmaker's focus.

Preparation can also help a candidate to stay in the running. Interviewees whose answers show that they are uninformed about their potential employer can expect to disqualify themselves. Richard Foust, a chemist at James Madison University in Harrisonburg, Virginia, points to his own place of work as an example. "What kills most candidates is, they don't understand that we're a predominantly undergraduate institution," he says. "Our first obligation is to fill the teaching position. Some candidates do the hard sell on their research — what they're working on as a postdoc. We weed those people out."

But to really get the dirt on what questions might be asked during an interview, candidates should try to obtain more information about the interviewer and their approach, ideally from current and former colleagues, mentors, advisers, supervisors or other trusted contacts. Juan Francisco Abenza Martínez says that he might have performed more effectively during a postdoctoral-research interview with a principal investigator a few years ago — or might have decided to scrap the interview altogether — had he known in advance about the person's rapid-fire interviewing style. Martínez, now a junior researcher in biophysics and genetics at the University of Cambridge, UK, was asked for an 'elevator presentation' — a ▶

five-minute explanation of his work. He blundered through his answer and didn't get the post. "I wasn't prepared for that," he says.

Debojyoti Dhar faced a similar curve ball in 2009 during an interview for a research post with a life-sciences company in India. It was his first experience with industry — fresh out of a postdoc at the University of Massachusetts Medical School in Worcester, Dhar was gobsmacked when asked if he could discover a drug target or vaccine and develop it in just six months.

"Scientists generally do not know the intricacies of business," he says. "I had no clue what to expect. I lost out." He later discovered that not all companies have such a "short-term-gain mentality", including the one for which he now works as vice-president, Leaf Cleantech in Bangalore.

STAYING COOL

No matter what the interview question, a calm, contemplative reply will win out almost every time, say hiring professionals. It reflects poise and an ability to maintain grace under pressure, instead of a panic to provide the 'best' answer.

But although being calm is paramount, memorizing replies for fear of losing one's cool or of giving the 'wrong' answer usually backfires, say interviewers. They can tell when an applicant is reciting — whether through notes (when interviewing by phone or video-conference) or from memory. "It's OK to be polished and practised, but I'm a person trying to have a conversation with you," says Jennifer Hobbs, director of training grant support and postdoctoral affairs at Northwestern University in Evanston, Illinois. "I need a sense of who you are."

"It's OK to be polished and practised, but I'm a person trying to have a conversation with you. I need a sense of who you are."
Jennifer Hobbs

Playing it safe doesn't win points either. Candidates who are relaxed and self-assured enough to step outside conventional interview protocol — whether by asking for more time or by turning the tables and asking the interviewer a question — are likely to stand a better chance of getting an offer.

Senior programme manager Marina Ramon recalls being pleasantly surprised by one applicant's quiet, composed response when asked how her short- and long-term goals aligned with the mission of the society she was seeking to work for. "She said, 'I can't answer this question immediately — I need time to think about it and to synthesize all the different elements that I want to incorporate'," says Ramon, who works at the Society for Advancement of Chicanos and Native Americans in Science in Santa Cruz, California. Ramon agreed that the candidate could e-mail a reply within several hours. The candidate did just that, and got the job.

"There are questions that require some thoughtful consideration. It is fair to ask for time to think about a question."
Marina Ramon

"It isn't that you have to think on the fly," says Ramon. "There are questions that require some thoughtful consideration, and whether you panic, or stand back and contemplate, can make a difference in whether you're offered the position. It is fair to ask for time to think about a question."

Ideally, candidates will remain self-possessed enough to stay a step ahead of the interviewer, a quality that sold Keith Micoli on a candidate for an assistant-director post at New York University (NYU). A job candidate turned the tables on Micoli by serenely posing a question that he had never thought to ask himself.

"She asked me what I would consider a successful first year for the person who was hired," says Micoli, director of the postdoctoral programme at NYU Langone Medical Center. Taken aback, he realized he had not decided exactly what the new hire would do. "It instantly made her the successful candidate in my mind." The question, to Micoli, reflected a sincere interest in the job and a desire to succeed. He hired her.

Since his own job interview back in 2011, Roman has himself interviewed many applicants for posts in the South African ministry. He recommends that candidates who are struggling with a tricky question ask the interviewer to repeat it, which helps to ensure that they had heard it correctly. The strategy also buys the candidate time to consider a thoughtful response. But he agrees that behaviour and demeanour are ultimately more significant than the answer itself.

"I look for someone who doesn't get flustered and who can remain calm," he says. "No matter what they say, if they can convey that they're at ease, confident and sure of themselves, it's all good." ■

Karen Kaplan *is associate Careers editor at* Nature.

THE 1-2-3 OF INTERVIEWS

Steps to success for researcher applicants.

1 THE ODD QUESTIONS

Here are possible ways to address tricky questions reported by early-career scientists:

- **What was your favourite story in today's *New York Times*?** If you don't actually read that newspaper, just say non-defensively that, as a scientist, you get ideas from many sources, and discuss a story or post you read that day. You could also ask the interviewer whether there was a particular story that they found interesting.
- **How would you describe yourself?** Such a vague, cryptic request can be puzzling. You can talk first about your work and research experience, but you should also bear in mind that this is a way of getting at who you are as a person. It can therefore be useful to explain how you work well as a team player — give examples of collaborations or other teamwork — and provide some personal information, such as hobbies, and how they relate to the job.
- **What is your three-year plan?** You should know from the job advert (and some background research) what the organization's mission is, so discuss how your plan aligns with that mission.
- **What do your parents do for a living?** You might ask the interviewer why he or she is asking the question as it is quite personal. The question might be a reflection of a poor or misinformed interviewer. Or it might be intended to see whether you can remain diplomatic and keep your head.

See more interview questions and ways to handle them at go.nature.com/elnmcp.

2 INTERVIEW INTELLIGENCE

Here are ways to find information about your potential employer and about the person or team who will interview you.

- Visit the employer's website (the university department, the principal investigator's laboratory page or the agency or company's website) for an overview.
- Look at business directories, scientific publications and newsletters. Read news sources to learn about research funding, corporate mergers and product launches, recommends Deb Koen, a career strategist and *Nature* Careers columnist in Rochester, New York.
- For anonymous employee reviews of companies and institutions, see sites such as www.glassdoor.com and www.jobitorial.com. But remember that postings are the opinions of individual employees and might not accurately reflect views overall. Compare this feedback with insights from other sources, says Koen.
- To get a sense of the organization's culture, examine its social-media presence, including on Twitter, Google+, Facebook and blogs, advises executive coach Louellen Essex in Minneapolis, Minnesota.
- Ask the recruiter or person who scheduled your interview (by phone or e-mail) to disclose who will be interviewing you and their positions. Look at their web pages, academic publications and social-media presence.
- Reach out to contacts from LinkedIn and other social media, as well as to graduate-school fellow alumni, former labmates and colleagues from scientific societies, for information about the organization and your interviewers.

3 STAYING COOL

Winning at the interview game often depends on staying composed and calm. Below are some suggestions for mitigating stress during the exchange.

- Look at careers websites for examples of difficult or puzzling questions that interviewers across all sectors have posed in the past, says Lee Miller, a career coach and columnist in New York City. Rehearse for the interview with a friend or colleague, and practise different ways to respond to those questions.
- Bring your CV and prepare a list of questions to ask the interviewer at the end of the discussion. Make notes to yourself on the list to breathe, slow down and pause, and refer to these notes during the interview.
- If you are completely in the dark about how to respond to a question, say: "Before I answer that, let me ask you this," and pull out a related question from your list. Or ask for more time or for the next question.
- Counter self-sabotaging thoughts, says Koen. Change "I'll never be prepared enough for this interview. It will be a disaster with questions I don't know how to answer" into "I am prepared for this interview. If I'm unsure of an answer, I will remain calm and make a positive impression overall."
- If you're not satisfied with one of your responses in the interview, you can re-address it in your follow-up thank-you letter or e-mail, notes Koen. K.K.

MENTORING

Balancing act

Just one-fifth of US clinician–researchers report receiving guidance from mentors on achieving work–life balance, finds a survey (R. DeCastro *et al. Acad. Med.* **89,** 301–311; 2014). The authors polled 1,227 researchers who received National Institutes of Health career-development grants in 2006–09. They found that although 52% of female respondents and 40% of male respondents were dissatisfied with their work–life balance, only 22% of all people surveyed received advice from a mentor on balancing the two. Researchers should not fear initiating discussions about such issues with their advisers, says co-author Reshma Jagsi, a radiation oncologist at the University of Michigan in Ann Arbor, who adds that mentors may not be aware of their mentees' work–life conflicts. "This is not an illegitimate concern," she says.

DEGREES

Enrolment slowdown

First-time enrolment in US professional master's degree (PSM) programmes continues to rise, but the rate of increase is slowing, finds a report from the Council of Graduate Schools (CGS) in Washington DC. Enrolment rose by 2.2% in 2012–13, compared with increases of 11.7% in 2011–12 and 14.7% in 2010–11. The drop corresponds to a slowdown in enrolment in all US graduate programmes, says Jeffrey Allum, director of research and policy analysis for the CGS. He notes that a 2013 CGS survey found that 91% of PSM graduates were working in their field of study and 68% of full-time employees had annual earnings above US$50,000.

TRAINING

Support means success

Junior researchers need stronger career-development support and training, says *The Global State of Young Scientists,* an analysis by the Global Young Academy in Berlin. It surveyed 650 early-career researchers aged 30–40 worldwide in 2013. Respondents said that solid mentoring relationships are vital for career success, in part by providing access to research groups and opportunities for giving talks and publishing papers. But many respondents described existing adviser support as inadequate. Co-author Catherine Beaudry, associate professor of innovation economics at the Polytechnic School of Montreal, Canada, counsels researchers to seek support from many senior colleagues.

COFFEE IN END TIMES

Grounds for despair.

BY ALVARO ZINOS-AMARO & ALEX SHVARTSMAN

Henry Lindnis measured out his life with coffee spoons.

It started when Henry and his wife, Erin, watched the President deliver the news on their TV set.

Erin's hand found Henry's as they listened to the President's sombre voice. He announced that a fleet of huge alien vessels had been detected approaching the Solar System. They would arrive on Earth in four weeks. He urged everyone to remain calm.

"I'm going to make us espresso," said Henry.

Espresso was something they reserved for special occasions.

"You know, hon, I've given up coffee," Erin said, without making eye contact.

"You've what?"

"I told you I was going to. Last week. When we were having dinner with the Tillmans. Pat made a great point about how acidic it is."

Henry didn't recall the conversation. "You're worried about the health effects of coffee when we have a month left to live?"

Erin gasped. "You think it's an invasion?"

"If they meant to explore or make contact, the aliens would send a single ship. This is a war armada." He pointed at the grainy video on the screen, sent back by the probes from the far edge of the Solar System.

"Always the pessimist."

"Even if I'm wrong, things are going to get bad. People won't react well. We need to get out of the city right away. We can go to my dad's old cabin." He disengaged his hand from Erin's. Just because she chose the worst possible time to quit caffeine didn't mean he had to deprive himself. "I'm having an espresso. Then we're going to pack."

Images of the alien ships filled the screen. They were ugly things, unwieldy, grotesquely large, the size of islands, shapes with jagged metal protrusions and incomprehensible ports and turrets and antennas, like insects made of polymer and metal. Dark grey against the black of space. And they were getting closer. Every minute. Every second.

NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

Erin went along grudgingly, complaining and calling Henry paranoid. But she helped him pack. They loaded the truck with supplies and drove to the ramshackle Northern California cabin.

He felt almost vindicated when the reports began to come in over the radio.

There were widespread riots and looting. Supermarkets and warehouses were cleaned out of all essentials, and people fought for whatever remained. The President declared a state of emergency and called in the army reserves. And that was only the beginning.

JACEY

Henry and Erin spent three weeks in the cabin, with nothing but a few books and a radio. During the day Henry kept busy, fishing and hunting. In the evenings, they talked. Or tried to. When had their conversations become so forced?

He made himself two cups of coffee per day, rationing their last container of ground beans. It would have only been one cup per day for each of them, but Erin refused to have any.

If we can't share a cup of coffee, what can we share? Henry thought. After the first week, they hardly talked at all.

One evening, Henry heard motor sounds. He looked out the window and saw a Ford pickup pull up to the cabin. A man and a woman stepped out of the truck.

He walked out of the cabin, rifle in hand. "You're trespassing," he said.

"We don't want any trouble," said the man. "We're just looking for a safe place. It's dangerous out there. Can we at least stay the night? It's getting dark."

A pair of small children, eyes wide, stared at him from the Ford's back seat.

Henry lifted his weapon. "Look somewhere else."

The intruders scrambled back into the Ford.

"I can't believe you did that," Erin called out from the doorway.

"We have limited resources, not enough to share with strangers," said Henry.

"The Henry I married would never have turned away scared children," said Erin. "You've changed."

"It's the world that's changed, Erin. You have to get used to that."

In the morning she packed a bag of clothes. "I'm driving home."

He didn't try to talk her out of it.

Henry ran out of coffee the day before the aliens were due to reach Earth. He brewed the last cup, savoured its fine aroma, and set it down on the nightstable.

Right next to the pistol.

Coffee had taught him many things over the years, and even this last lesson in moderation had proved interesting.

Interesting, but ultimately pointless.

Just like everything would be in the aftermath of the alien apocalypse. With no coffee. And no Erin.

Henry exhaled loudly and let go of his memories, his sense of self, his awareness of anything except for the cup and the gun. Then he drank the warm brown liquid in short sips, and smiled.

A minute later, less than a mile away, an elk was momentarily startled at the sound of a gunshot.

The next day, the aliens arrived bearing gifts: the ability to bend space and time, groundbreaking mathematical theorems, stable cold fusion.

And a roast that was *inhumanly* tasty. ■

Alvaro Zinos-Amaro *and* **Alex Shvartsman** *are American science-fiction writers. Learn more about them at myaineko.blogspot.com and alexshvartsman.com.*